Departamento de Ciências e Tecnologias

Mestrado em Estatística, Matemática e Computação

Teste *F* na Regressão Linear Múltipla para Dados Temporais com Correlação Serial.

Bruno Fernando Pinheiro Faria

Lisboa, 2011

Mestrado em Estatística, Matemática e Computação

Teste *F* na Regressão Linear Múltipla para Dados Temporais com Correlação Serial.

Bruno Fernando Pinheiro Faria

Dissertação apresentada para obtenção de Grau de Mestre em
Estatística, Matemática e Computação, especialidade Estatística Computacional

Orientadora: Professora Doutora Maria do Rosário Ramos

(Professora Auxiliar da Universidade Aberta)

Lisboa, 2011

Resumo

Na análise de regressão linear, simples ou múltipla, o teste *F* é utilizado para testar simultaneamente a significância de um conjunto ou um subconjunto de parâmetros. Neste trabalho, é estudado o comportamento da estatística F usual para testar a significância dos coeficientes sazonais no modelo de regressão linear múltipla para séries temporais com tendência, sazonalidade e correlação serial.

Quando alguns dos pressupostos de validade do teste são violados é de se esperar que o teste seja afetado. Por isso, analisa-se, através de um estudo de simulação de Monte Carlo, o comportamento da estatística *F* quando são violados os pressupostos de normalidade e da independência dos erros num caso específico de modelo de autocorrelação – AR(1).

O estudo de simulação para avaliar a *performance* do teste *F* usual foi realizado sob vários cenários, tendo em conta desvios da normalidade e considerando que os erros do modelo são correlacionados, com diferentes intensidades da autocorrelação. O comportamento dos testes é avaliado com base nos valores do nível de significância empírico e da potência dos testes.

Além disso, apresenta-se teoricamente a estatística de teste F para casos em que a autocorrelação nos erros é considerada através da estimação pelos mínimos quadrados generalizados e realiza-se, através de simulação, um estudo comparativo do comportamento deste teste, quando a autocorrelação é incorretamente estimada.

Finalmente são apresentadas duas aplicações em séries temporais, uma respeitante ao número de hóspedes em hotéis tradicionais e casas de turismo rural nos Açores e a outra relativa ao número de exportações de automóveis em Portugal.

Palavras-Chave: Séries Temporais, Modelo Sazonal, Teste F, Regressão Linear Múltipla, Modelo Autorregressivo.

Abstract

In simple or multiple linear regression analysis, the F test is used to test simultaneously the significance of a set or a subset of parameters. In this assignment it is studied the behaviour of the usual F statistic to test the significance of seasonal coefficients in multiple linear regression model for time series with trend, seasonality

When some of the test assumptions are violated, we have to assume that the test can be affected. Therefore, we will analyse, through a Monte Carlo simulation study, the behaviour of the F statistic, when the normality assumptions and the independence of the error term in a specific case of autocorrelation model AR(1) are violated.

The simulation study to assess the performance of the usual F test was made under several situations, bearing in mind normality deviations and considering that the errors of the autoregressive model are correlated with different autocorrelation levels. The behaviour of the tests is evaluated through the empirical significance level and the power of the test.

Furthermore, we present theoretically the F statistic in cases in which the autocorrelation of the error term is taken into account through general least squares estimation and we make, through a simulation study, a comparative analysis of the aforementioned test behaviour, when the autocorrelation of the error term is incorrectly estimated.

Finally, we will show two applications in a time series, one regarding the number of guests in traditional hotels and rural tourism in Azores and the other the number of cars exportations in Portugal.

Keywords

Time Series, Seasonal Model, F Test, Multiple Linear Regression, Autoregressive Model

Agradecimentos

Agradeço à Professora Doutora Maria do Rosário Ramos a orientação, o apoio e a disponibilidade com que me acompanhou ao longo da elaboração desta dissertação.

Agradeço à minha namorada pelo auxílio prestado na tradução em alguns dos trabalhos elaborados neste mestrado, pela compreensão e paciência dispensada ao longo destes dois últimos anos, por tudo o que deixei de fazer em virtude do trabalho e pela forma que sempre me apoiou, principalmente nos momentos mais difíceis.

Agradeço também à minha família pelo apoio dado durante todo o meu percurso académico e que sem ele eu não teria chegado até aqui.

Índice

Índice de Tabelas

Índice de Figuras

1. Introdução

A sazonalidade está muitas vezes presente em variáveis definidas no tempo, quando a ordem de ocorrência é relevante, sendo exemplos as áreas do turismo, economia ou biologia. Devido à importância que esta componente normalmente tem no estudo destas variáveis, é importante que um modelo matemático que seja usado no ajustamento destas séries temporais reflita a variação atribuída à sazonalidade, caso esta seja significativa. Para testar a significância desta componente, no caso específico do modelo com tendência linear, é normalmente utilizada a estatística de teste F.

O problema principal que nos propomos aprofundar é o da performance do teste F usual na análise da sazonalidade numa série temporal com tendência, em algumas situações específicas que podem ou não ocorrer simultaneamente, nomeadamente:

- Quando as observações são dependentes, ou seja, quando existe alguma estrutura de autocorrelação na série – neste caso, quando a série dos erros é gerada por um processo AR(1);
- Quando a distribuição dos valores da série se afasta da distribuição normal, apresentando vários níveis de assimetria.

Neste trabalho, é realizada uma primeira abordagem teórica à temática estudada, com base na literatura existente, onde é apresentado o modelo linear de séries temporais com tendência e sazonalidade, assim como a expressão da estatística de teste F em modelos completos e reduzidos de regressão linear múltipla quando a estimação dos coeficientes do modelo é estimada pelo método dos mínimos quadrados, quer ordinários, quer generalizados. Faz-se ainda uma análise aos erros associados ao modelo, com incidência nos processos autorregressivos de 1ª ordem estacionários.

Para completar o desenvolvimento teórico é realizado um estudo de simulação estocástica, com o objetivo de comparar o comportamento do teste F segundo variações dos parâmetros iniciais da série temporal simulada e obter informações conclusivas acerca das seguintes questões:

- Como é afetado o nível de significância empírico e a potência empírica do teste F para cada uma das situações acima descritas?
- Qual a consequência da estimação do parâmetro de autocorrelação no teste F?

1. Introdução

Este trabalho desenvolve-se ao longo de 10 capítulos que podem ser organizados em duas partes mais abrangentes. A primeira parte engloba os capítulos 2, 3 e 4 onde é efetuado o enquadramento do problema e apresentado o desenvolvimento teórico e a descrição dos métodos usados ao longo do trabalho. A segunda parte reúne os capítulos 5, 6 e 7 e compreende um estudo de simulação para avaliar o comportamento do teste F segundo diferentes parâmetros iniciais, duas aplicações a séries de variáveis da área do turismo e da indústria, respetivamente, bem como as conclusões e os comentários finais.

No capítulo 2 apresenta-se formalmente o modelo linear aditivo representativo das séries temporais com tendência e sazonalidade a partir do qual se desenvolve todo o estudo subsequente.

No capítulo 3 é apresentado o método de estimação pelos mínimos quadrados ordinários e a expressão da estatística F no teste à significância de um modelo linear (completo ou reduzido). É definido formalmente o teste de hipóteses respeitante ao problema principal deste trabalho, é apresentada a estatística F para testar a significância da componente sazonal numa série temporal com tendência e sazonalidade e é feita uma breve referência aos erros de tipo I e de tipo II, inerentes ao teste de hipóteses definido.

No capítulo 4 é abordado o processo autorregressivo de 1ª ordem, através do qual é definida a componente estocástica das séries temporais em estudo neste trabalho. É realizada uma caracterização do processo e são definidas as funções de autocorrelação e de autocorrelação parcial, através das quais este processo pode ser identificado por análise à componente dos resíduos.

No capítulo 5 é apresentada a expressão da estatística *F* em que a autocorrelação nos erros é tida em conta através da estimação pelos mínimos quadrados generalizados. É abordado, de forma resumida, o método de Cochrane e Orkut, para estimação dos coeficientes do modelo em estudo quando os erros apresentam autocorrelação, e são referenciados alguns resultados de estudos já realizados.

No capítulo 6 é apresentado um estudo de simulação para averiguar a *performance* da estatística *F* usual, no teste à significância da componente sazonal, em séries em que a componente dos erros apresenta correlação, nomeadamente, quando esta é um processo autorregressivo de 1ª ordem. A análise é feita tendo em conta o nível de significância empírico e a potência empírica do teste.

São experimentadas várias situações que podem ter influência no comportamento do teste, nomeadamente a dimensão da amostra simulada, o nível de autocorrelação da série dos erros e o desvio à normalidade, com a inclusão da distribuição gama na componente estocástica.

É também analisada a influência do parâmetro de tendência e da constante β_0 na estatística F, o comportamento do teste F quando a autocorrelação é tomada em linha de conta através da estimação pelos mínimos quadrados generalizados, em particular, quando a estimativa do parâmetro autorregressivo afasta-se do verdadeiro valor populacional.

No capítulo 7 são apresentadas duas aplicações de análise ao comportamento da estatística *F* usual no teste à significância da componente sazonal, uma relativa à variável número de dormidas em espaços de turismo rural e hotelaria tradicional nos Açores e outra referente à variável exportações de automóveis em Portugal. No primeiro caso, a sazonalidade está claramente definida, fato este comprovado pela aplicação do teste F, enquanto, na segunda série, a aplicação do teste *F* usual revela-se inconclusiva.

Por último, no capítulo 8 é apresentado um resumo dos resultados mais importantes e são sugeridas ideias para investigações futuras.

2. Séries Temporais com Tendência e Sazonalidade

2.1. Séries Temporais

Uma série temporal pode ser definida como uma sucessão de observações ordenadas no tempo. São exemplos a precepitação diária, o número de nascimentos numa dada região ou o volume de vendas anuais de uma empresa.

Em estatística, assim como noutros campos da investigação, como a econometria, a matemática financeira e a bioestatística, é usual estudarem-se séries temporais que apresentam algum tipo de variabilidade, à qual estão normalmente associadas perturbações aleatórias. O estudo destas sequências permite a caracterização dos processos por estas representadas, a previsão do comportamento destes processos em termos futuros e, por último, a identificação e avaliação dos fatores que possam ter influenciado o comportamento das mesmas.

Dada a natureza dos valores das séries temporais, estas podem ser classificadas como discretas ou contínuas. Uma série temporal diz-se discreta quando apenas pode ser observada em intervalos de tempo regulares. São exemplos o número de passageiros mensais de uma companhia aérea, o volume de vendas anual e os lucros trimestrais. Nos casos em que os valores podem ser registados em qualquer momento temporal, sem interrupções, as séries dizem-se contínuas. Como exemplos temos a velocidade do vento, a pressão e a temperatura. Neste trabalho são consideradas apenas séries discretas ou séries originalmente contínuas em que registos são realizados em intervalos de tempo regulares.

2.2. Modelos com Tendência e Sazonalidade

Para caracterizar e estudar uma série temporal é fundamental encontrar um modelo matemático que se aproxime dos valores em estudo. Normalmente as séries temporais são analisadas tendo em conta aspetos como a tendência, o ciclo, a sazonalidade e as variações aleatórias.

Um dos modelos usados no estudo destas séries é um modelo com tendência simples, caracterizado pela seguinte expressão:

$$Z_t = \mu_t + \varepsilon_t, \quad t = 1 \leq t \leq N$$

em que os erros aleatórios, ε_t, são tais que $E(\varepsilon_t) = 0 \quad \text{var}(\varepsilon_t) = \sigma^2$

Neste modelo, os valores observados dependem da componente determinística μ_t e da componente aleatória ε_t. Tem-se ainda que $E(Z_t) = \mu_t$.

Em muitas situações, o valor médio das séries temporais incorpora componentes que são funções com determinado tipo de variação. Em particular, um dos tipos de variação que ocorre frequentemente é o caso em que a série apresenta uma tendência linear. Neste caso, a série é caracterizada por,

$$Z_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Contudo, este modelo de tendência simples poderá não ser suficiente para caracterizar determinadas séries, uma vez que o seu valor poderá incorporar componentes que são funções com outro tipo de variação, tais como as periódicas.

Um modelo que incorpore, para além da tendência, uma variabilidade periódica designa-se de modelo linear com tendência e sazonalidade e pode ser caracterizado por um modelo aditivo com a seguinte expressão:

$$Z_t = \beta_0 + \beta_1 t + s_t + \varepsilon_t, \quad (2.2.1)$$

em que $\beta_0 + \beta_1 t$ representa a tendência, s_t representa a componente sazonal e ε_t representa a componente aleatória associada ao modelo.

Visto s_t representar uma função periódica de período P, tem-se que $s_{t+P} = s_t$. Os valores de $s_1, s_2, ..., s_P$ designam-se de coeficiente sazonais.

Dado que, do ponto de vista gráfico, β_0 representa a ordenada na origem, exige-se que $\sum_{t=1}^{P} s_t = 0$

Para facilitar a estimação dos coeficientes sazonais de uma série temporal com tendência a sazonalidade, recorre-se a uma transformação da equação (2.2.1). É então

definido um conjunto de variáveis auxiliares, chamadas indicatrizes, x_{ti}, de forma a que:

$$Z_t = \beta_0 + \beta_1 t + \sum_{i=1}^{P} x_{ti} s_i + \varepsilon_t$$

em que os s_i representam os coeficientes sazonais;

$$x_{ti} = \begin{cases} 1 \quad se \quad t = (j-1)P + i, \text{para algum } j = 1,..., N_i \\ 0, \quad \text{caso contrário} \end{cases}.$$

O facto de se ter considerado $\sum_{t=1}^{P} s_i = 0$ tem como consequência uma dependência linear entre as variáveis x_{ti}, $i = 1,..., P$. Esta condicionante fará com que a estimação dos parâmetros através dos mínimos quadrados não seja possível. Para contornar este problema, pode reescrever-se o modelo anterior na forma:

$$Z_t = \beta_1 t + \sum_{i=1}^{P} x_{ti} S_i + \varepsilon_t,$$

com $s_i = S_i - \overline{S}$

$\beta_0 = \overline{S}$

Tal como acontece com outros estudos em estatística, o ajustamento de um conjunto de dados a um modelo pressupõe a estimação de parâmetros populacionais, neste caso, dos coeficientes do modelo. Como se pretende um ajuste dos dados a um modelo linear, com mais do que uma variável independente, os valores esperados da variável dependente serão obtidos através da chamada regressão linear múltipla.

Existe mais do que um método para estimar os coeficientes de um modelo de regressão linear, contudo o método dos mínimos quadrados é o mais utilizado, pois, sob a hipótese de independência, normalidade e homocedasticidade dos resíduos, fornece estimadores centrados com variância mínima. O estudo do processo de regressão linear múltipla, com estimação pelos mínimos quadrados será abordado no próximo capítulo.

3. O Teste F sobre o Modelo de Regressão Linear Múltipla

3.1. O Modelo de Regressão Linear Múltipla

Nas séries temporais, assim como em muitos problemas de Estatística, que envolvem mais do que uma variável, é possível estabelecer algum tipo de relação entre as variáveis independentes e a variável dependente. No caso das séries com tendência linear e sazonalidade, estudadas neste trabalho, os dados são ajustados a um modelo através de regressão linear. Para modelos apenas com tendência, é aplicada a regressão linear simples, pois existe somente uma variável independente, o tempo. Por sua vez, quando é incorporada a componente de sazonalidade no modelo, é utilizada a regressão linear múltipla. De referir que as perturbações aleatórias inerentes às observações refletir-se-ão no modelo na chamada componente estocástica (dos erros aleatórios).

Na regressão linear múltipla assume-se uma relação linear entre uma variável aleatória *y*, e *k* variáveis independentes (ou regressores), $x_1, x_2, ..., x_k$. Assim sendo e, considerando os parâmetros $\beta_0, \beta_1, \beta_2, ..., \beta_p$, o modelo de regressão linear múltipla é definido através da seguinte expressão:

$$y_i = \beta_0 +, \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_k x_{k(i)} + \varepsilon_i \quad , \forall i = 1, ..., n \quad (3.1.1)$$

em que os parâmetros $\beta_j, j = 1, ..., p$ são chamados coeficientes de regressão;

e ε_i representam os erros aleatórios.

Neste modelo, os parâmetros β_j representam a variação esperada em *y* por unidade de variação em x_j, quando todas as restantes variáveis x_i, com $i \neq j$, permanecem constantes.

Na regressão linear múltipla assume-se os seguintes pressupostos, relativamente à componente dos erros aleatórios:

- $\varepsilon_i \quad \cap N(0, \sigma^2), \quad \forall i = 1, ..., n$.
- $\{\varepsilon_i\}_{i=1}^n$ variáveis aleatórias independentes.

A equação apresentada em (3.1.1) refere-se a um elemento genérico *i*. A relação entre as *n* observações é dada pelo seguinte conjunto de equações:

$$
\begin{aligned}
y_1 &= \beta_0 +, \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + ... + \beta_k x_{k(1)} + \varepsilon_1 \\
y_2 &= \beta_0 +, \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + ... + \beta_k x_{k(2)} + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad \vdots \qquad\qquad \vdots \\
y_n &= \beta_0 +, \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + ... + \beta_k x_{k(n)} + \varepsilon_n
\end{aligned}
$$

De forma a simplificar a escrita, pode-se recorrer à notação matricial para representar as equações anteriores. Assim sendo e, considerando as seguintes matrizes representativas de cada uma das variáveis em questão,

$$
Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \quad X = \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{k(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{k(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{k(n)} \end{bmatrix}_{n\times(k+1)}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}
$$

tem-se a seguinte equação matricial:

$$Y = X\beta + \varepsilon,$$

em que $\varepsilon_i \ \cap N(0, \sigma^2 \cdot I_n)$

Sublinhe-se que este tipo de representação, para além de simplificar a escrita, facilita também a implementação de procedimentos de cálculo, nomeadamente do ponto de vista computacional.

3.2. Estimação dos Parâmetros no Modelo de Regressão Linear Múltipla

No modelo de regressão linear múltipla, os parâmetros $\beta_k's$, referentes à população, são à partida desconhecidos, pelo que é necessário proceder-se à sua estimação, a partir de um conjunto de *n* observações dessa população. O método dos mínimos quadrados ordinários (MQO), proposto por Karl Gauss (1777 – 1855), é uma das técnicas mais utilizadas para estimar os parâmetros de um modelo linear, através da minimização da soma dos quadrados dos resíduos e_i (diferença entre os valores estimados e os dados observados).

Tomando $b^T = (b_0, b_1, ..., b_k)$ como o vetor da estimativa dos coeficientes de regressão, a estimativa da variável resposta é calculada a partir dos dados amostrais a partir da expressão

$$\hat{\mathbf{Y}} = \mathbf{Xb},$$

com $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$

Pretende-se então encontrar os estimadores de $\beta_k's$ que minimizem a quantidade $\sum_{i=1}^{n} e_i^2$, que na forma matricial pode ser escrita como $\mathbf{e^T e}$.

Utilizando a igualdade $\mathbf{e = Y - Xb}$, tem-se que

$$\begin{aligned}\mathbf{e^T e} &= \left(\mathbf{Y - Xb}\right)^{\mathbf{T}} \cdot \left(\mathbf{Y - Xb}\right) \\ &= \mathbf{Y^T Y - 2bb^T Y + b^T X^T Xb} \qquad (3.2.1)\end{aligned}$$

O passo seguinte será o de diferenciar a expressão (3.2.1) em ordem a b e igualar a zero, ou seja, resolver a seguinte equação:

$$\frac{\partial}{\partial \mathbf{b}}\left[\left(\mathbf{Y - Xb}\right)^{\mathbf{T}}\left(\mathbf{Y - Xb}\right)\right] = \mathbf{0}$$

que é equivalente a

$$\frac{\partial}{\partial \mathbf{b}}\left(\mathbf{Y^T Y - 2bX^T Y + b^T X^T Xb}\right) = \mathbf{0}\,.$$

Resolvendo a derivada parcial em relação a **b**, obtém-se a seguinte equação:

$$\mathbf{X^T Xb = X^T Y}\,.$$

Assim, desde que a expressão $\mathbf{X^T X}$ seja invertível, o estimador **b** é dado por:

$$\mathbf{b} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y} \qquad (3.2.2)$$

Se ocorrer colinearidade na matriz **X**, ou seja, se alguma variável x_k puder ser expressa por uma combinação linear dos restantes vetores coluna de **X**, a matriz $\left(\mathbf{X^T X}\right)^{-1}$ é singular, logo, não invertível. Para contornar estre problema, pode-se reparametrizar o modelo de modo a que as variáveis independentes que nele figurem sejam linearmente independentes e então aplicar a estimação pelos MQO.

Teorema 3.2.3: O estimador $\mathbf{b} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}$ é um estimador centrado e a sua variância é $\text{var}(\mathbf{b}) = \left(\mathbf{X^T X}\right)^{-1}\sigma^2$

Demonstração:

$$\begin{aligned}E(\mathbf{b}) &= E\left(\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right) = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}E(\mathbf{Y}) = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}E\left(\mathbf{X\beta + \varepsilon}\right) \\ &= \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T X\beta} = \boldsymbol{\beta}\end{aligned}$$

$$\text{var}(b) = \text{var}\left[\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right] = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\left[\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right]^{\mathbf{T}}\text{var}\mathbf{(Y)}$$

$$
\begin{aligned}
&= \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\left[\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right]^{\mathbf{T}} \text{var}(\mathbf{X\beta}+\boldsymbol{\varepsilon}) \\
&= \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\left[\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right]^{\mathbf{T}} \text{var}(\boldsymbol{\varepsilon}) \\
&= \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T X}\left[\left(\mathbf{X^T X}\right)^{-1}\right]^{\mathbf{T}}\sigma^2 \\
&= \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T X}\left(\mathbf{X^T X}\right)^{-1}\sigma^2 \\
&= \left(\mathbf{X^T X}\right)^{-1}\sigma^2 \quad \blacksquare
\end{aligned}
$$

O estimador *b* segue uma distribuição $b \cap N\left(\boldsymbol{\beta}, \left(\mathbf{X^T X}\right)^{-1}\sigma^2\right)$.

Teorema 3.2.4. (Gauss-Markov) – Dadas as hipóteses do modelo clássico de regressão linear, o estimador pelos mínimos quadrados ordinários, é o mais eficiente entre todos os estimadores lineares não enviesados, pois apresenta variância mínima.

Teorema 3.2.5: O estimador centrado para a variância σ^2 dos erros aleatórios é dado por:

$$s^2 = \frac{\mathbf{e^T e}}{n-k-1} = \frac{\mathbf{Y^T}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}}{n-k-1} = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

Demonstração:

A soma dos quadrados dos resíduos na regressão linear múltipla, dada por $e^T e$, pode ser escrita como:

$$
\begin{aligned}
\mathbf{e^T e} &= \left(\mathbf{Y}-\mathbf{Xb}\right)^{\mathbf{T}}\left(\mathbf{Y}-\mathbf{Xb}\right) \\
&= \left(\mathbf{Y}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right)^{\mathbf{T}}\left(\mathbf{Y}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right) \\
&= \mathbf{Y^T Y}-\mathbf{Y^T X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}-\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right)^{\mathbf{T}}\mathbf{Y}+\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right)^{\mathbf{T}}\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y} \\
&= \mathbf{Y^T Y}-\mathbf{Y^T X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}-\mathbf{Y^T X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}+\mathbf{Y^T X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y} \\
&= \mathbf{Y^T}\left(\mathbf{Y}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}\right) \\
&= \mathbf{Y^T}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}
\end{aligned}
$$

Esta última expressão corresponde a uma forma quadrática de variáveis aleatórias, do tipo $\mathbf{Y^T P Y}$, em que $\mathbf{Y}$ segue uma distribuição multivariada normal, $\mathbf{P} = \mathbf{I}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}$ e cujo valor esperado é $E\left(\mathbf{Y^T P Y}\right) = \sigma^2 tr(\mathbf{P}) + \boldsymbol{\mu}^{\mathbf{T}}\mathbf{P}\boldsymbol{\mu}$, onde $\boldsymbol{\mu}$ é o vetor média de $\mathbf{Y}$.

Tem-se ainda que

$$
\begin{aligned}
tr\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right) &= tr\left(\mathbf{I_n}\right)-tr\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right) \\
&= tr\left(\mathbf{I}_n\right)-tr\left(\mathbf{X^T X}\left(\mathbf{X^T X}\right)^{-1}\right) \\
&= tr\left(\mathbf{I}_n\right)-tr\left(\mathbf{I}_{k+1}\right) = n-k-1.
\end{aligned}
$$

Então, o valor esperado da soma dos quadrados dos resíduos é dado por

$$\begin{aligned} E(\mathbf{e}^{\mathbf{T}}\mathbf{e}) &= E\left(\mathbf{Y}^{\mathbf{T}}\left(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\right)\mathbf{Y}\right) \\ &= \sigma^2 tr\left(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\right)+(\mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}\left(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\right)(\mathbf{X}\boldsymbol{\beta}) \\ &= \sigma^2(n-k-1)+(\mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}\left(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\right)(\mathbf{X}\boldsymbol{\beta}) \\ &= \sigma^2(n-k-1)+(\mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}\left(\mathbf{X}\boldsymbol{\beta}-\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}\boldsymbol{\beta}\right) \\ &= \sigma^2(n-k-1). \end{aligned}$$

A partir daqui deduz-se o estimador centrado σ^2:

$$E(s^2)=\frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{n-k-1}=\frac{\sigma^2(n-k-1)}{n-k-1}=\sigma^2 \quad \blacksquare$$

3.3. O Teste F de Ajustamento Global do Modelo de Regressão Linear Múltipla

A estatística *F* usual, proposta por Snedcoor, testa a significância de um conjunto de parâmetros de um modelo de regressão linear múltipla, quando o modelo é ajustado aos dados através do método dos mínimos quadrados.

Define-se o seguinte teste de hipóteses:

$$H_0: \beta_0=\beta_1=\beta_2=...=\beta_k=0 \quad \text{vs} \quad H_1: \exists j: \beta_j \neq 0, \quad j=0,...,k$$

Utilizando a notação matricial, as hipóteses a testar assumem o seguinte aspeto:

$$H_0: \boldsymbol{\beta}=\mathbf{0} \quad \text{vs} \quad H_1: \boldsymbol{\beta}\neq\mathbf{0}$$

A variabilidade total dos valores da variável dependente Y, expressa através da soma dos quadrados dos desvios de Y face ao seu valor médio $\overline{Y}$ (SQT), pode ser separada em duas componentes aditivas: uma explicada pelo modelo de regressão (SQR) e a outra atribuída aos resíduos (SQE).

Considerando então as quantidades

	Notação matricial
$SQE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$SQE=(\mathbf{Y}-\mathbf{Xb})^{\mathbf{T}}\cdot(\mathbf{Y}-\mathbf{Xb})=\mathbf{e}^{\mathbf{T}}\mathbf{e}$
$SQR=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$	$SQR=(\mathbf{Xb}-\overline{\mathbf{Y}})^{\mathbf{T}}\cdot(\mathbf{Xb}-\overline{\mathbf{Y}})$
$SQT=\sum_{i=1}^{n}(y_i-\overline{y})^2$	$SQT=(\mathbf{Y}-\overline{\mathbf{Y}})^{\mathbf{T}}\cdot(\mathbf{Y}-\overline{\mathbf{Y}}),$

estabelece-se a seguinte relação:

$$SQT = SQR + SQE \text{ , ou seja,}$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

A fração da variância total em **Y** explicada pelo modelo de regressão e a da variância atribuída aos resíduos podem ser estimadas respetivamente pela divisão dos valores de *SQR* e *SQE* pelos seus graus de liberdade. Estas quantidades designam-se de quadrados médios da regressão (QMR) e quadrados médios dos resíduos (QME), respetivamente. Tem-se então que:

$$QM\text{R} = \frac{SQR}{k},$$

$$QME = \frac{SQE}{n\text{-}k\text{-}1} = \frac{\mathbf{e}^{\mathbf{t}}\mathbf{e}}{n-k-1} \ .$$

Para avaliar a qualidade do ajuste do modelo de regressão aos dados populacionais, compara-se a fração da variância explicada pelo modelo de regressão com a da variância atribuída aos resíduos. Assim, caso a primeira seja significativamente superior à segunda, podemos concluir que o modelo é significativo. Esta comparação é efetuada com base na distribuição estatística da razão entre estas duas variâncias.

Sob os pressupostos de independência e homocedasticidade dos erros aleatórios e sob a hipótese inicial H_0, tem-se que:

$$\frac{\text{SQR}}{k} \sim \sigma^2 \chi_k^2 \quad \text{e} \quad \frac{\text{SQE}}{n-k-1} \sim \sigma^2 \chi_{n-k-1}^2$$

Como estas duas quantidades são independentes, a estatística de teste, que corresponde à razão entre as duas quantidades anteriores, segue uma distribuição F com *k* e *n*-*k*-1 graus de liberdade, ou seja,

$$F = \frac{SQR/k}{SQE/n-k-1} \sim F_{k,n-k-1}$$

Na forma matricial, a estatística de teste F é dada por:

$$F = \frac{\mathbf{b}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b}/k}{\left[\mathbf{Y}^{\mathbf{T}}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\right)\mathbf{Y}\right]/n-k-1} \quad .$$

A partir das expressões anteriores podemos definir uma tabela de análise de variância do modelo ajustado de regressão linear múltipla.

Origem da variação	Graus de liberdade	Soma dos quadrados	Quadrados médios	Estatística de teste F
Regressão	*k*	SQR	$QM\text{R} = \frac{SQR}{k}$	$F = \frac{QM\text{R}}{QM\text{E}}$
Residual	*n-k-1*	SQE	$QME = \frac{SQE}{n\text{-}k\text{-}1}$	
Total	*n*-1	SQT		

Tabela 2. Análise de Variância para o teste F sobre o completo reduzido de regressão linear múltipla

múltipla A estatística de teste F, sob H_0, segue uma distribuição $F_{k,n-k-1}$. A um nível de significância α, rejeita-se H_0 se $F > F^{-1}{}_{k,n-k-1}(1-\alpha)$.

Sob a veracidade da hipótese alternativa $H_1 : \boldsymbol{\beta_1} \neq \mathbf{0}$, a estatística de teste *F* segue uma distribuição *F* não central com k e $n-k-1$ graus de liberdade e parâmetro de não centralidade dado por

$$\delta = \frac{\beta^T X^T X \beta}{\sigma^2} .$$

3.4. Teste F de Ajustamento para o Modelo Reduzido

A aplicação do teste F ao modelo completo de regressão linear resulta numa decisão sobre o conjunto de todos os parâmetros envolvidos no modelo. Portanto, uma rejeição da hipótese nula expressa que pelo menos um dos parâmetros β_j é significativamente diferente de zero, embora isto não signifique que todos o sejam.

Para se averiguar se um subconjunto de parâmetros é significativo no modelo de regressão linear, é necessário considerar-se um submodelo do modelo de regressão linear múltipla, de onde são excluídos os preditores correspondentes a esses parâmetros. Pretende-se com este procedimento apurar se as variáveis que foram excluídas do modelo reduzido são ou não significativas para o ajustamento global.

Considere-se então as partições $\mathbf{X_1}$ e $\mathbf{X_2}$, da matriz **X**, de modo a que $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$,

em que $\mathbf{X_1}$ é uma matriz de dimensão $n \times k_1$ e $\mathbf{X_2}$ é uma matriz de dimensão $n \times k_2$.

Pode-se reescrever o modelo completo $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ como:

$$\mathbf{Y} = \mathbf{X_1\beta_1} + \mathbf{X_2\beta_2} + \boldsymbol{\varepsilon}, \text{ com } \mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$$

O modelo reduzido de regressão linear é então definido por

$$\mathbf{Y} = \mathbf{X_2\beta_2} + \boldsymbol{\varepsilon}$$

Pretende-se averiguar se a inclusão de X_1 é significativa para o modelo, o que pode ser expresso pelo seguinte teste de hipóteses:

$$H_0 : \boldsymbol{\beta_1} = \mathbf{0} \quad vs \quad \mathbf{H_1 : \boldsymbol{\beta_1} \neq 0}$$

Tal como no modelo completo de regressão linear, a significância de um subconjunto de parâmetros pode ser aferida a partir da fração da variabilidade total explicada pela regressão, neste caso $\mathrm{SQR}_{\mathrm{completo}} - SQR_{reduzido}$, e da atribuída aos resíduos, $\mathrm{SQE}_{completo}$. De forma equivalente, pode-se escrever $\mathrm{SQR}_{\mathrm{completo}} - SQR_{reduzido}$ como $\mathrm{SQE}_{\mathrm{Reduzido}} - \mathrm{SQE}_{\mathrm{completo}}$.

Sob a hipótese inicial H_0, as quantidades $\mathrm{SQE}_{\mathrm{Reduzido}} - \mathrm{SQE}_{\mathrm{completo}}$ e $\mathrm{SQE}_{completo}$ divididas pelos respectivos graus de liberdades são duas variáveis aleatórias, tais que:

$$\frac{\mathrm{SQE}_{reduzido} - \mathrm{SQE}_{completo}}{k_1} \sim \sigma^2 \chi^2_{k_1} \quad \text{e} \quad \frac{\mathrm{SQE}_{completo}}{n-k-1} \sim \sigma^2 \chi^2_{n-k-1}.$$

Como estas duas variáveis aleatórias são independentes, a estatística de teste, definida como a razão entre essas duas quantidades, segue uma distribuição *F* com k_1 e n-k-1 graus de liberdade, ou seja,

$$F = \frac{\left(\mathrm{SQE}_{reduzido} - \mathrm{SQE}_{completo}\right)/k_1}{\mathrm{SQE}_{completo}/(n-k-1)} \sim F_{k_1, n-k-1}$$

A informação relativa ao modelo reduzido de regressão linear múltipla pode ser resumida no seguinte quadro de análise de variância:

Origem da variação	g.l	Soma dos quadrados	Quadrados médios	Estatística de teste F
Regressão	k_1	$\text{SQE}_{\text{reduzido}} - \text{SQE}_{completo}$	$QM\text{R}^* = \frac{\text{SQE}_{reduzido} - SQE_{completo}}{k_1}$	$F = \frac{\text{QMR}^*}{\text{QME}_{completo}}$
Residual	n-k-1	$\text{SQE}_{\text{completo}}$	$QM\text{E} = \frac{\text{SQRes}}{n\text{-}k\text{-}1}$	
Total	n-1	SQT		

Tabela 2. Análise de Variância para o teste F sobre o modelo reduzido de regressão linear múltipla

Usando a notação matricial, tem-se que:

$$SQE_{reduzido} = \mathbf{Y^T}\left(\mathbf{I} - \mathbf{X_2}\left(\mathbf{X_2^T X_2}\right)^{-1}\mathbf{X_2^T}\right)\mathbf{Y},$$

$$SQE_{completo} = \mathbf{Y^T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^{T'}}\right)\mathbf{Y},$$

$$SQE_{reduzido} - SQE_{completo} = \mathbf{Y^T}\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T I} - \mathbf{X_2}\left(\mathbf{X_2^T X_2}\right)^{-1}\mathbf{X_2^T}\right)\mathbf{Y}.$$

Logo, a estatística de teste F pode ser escrita através da seguinte razão:

$$F = \frac{\left[\mathbf{Y^T}\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T I} - \mathbf{X_2}\left(\mathbf{X_2^T X_2}\right)^{-1}\mathbf{X_2^T}\right)\mathbf{Y}\right] \Big/ k_1}{\left[\mathbf{Y^T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}\right] \Big/ (n-k-1)} \sim F_{k_1, n-k-1} \quad .$$

Sob a veracidade da hipótese alternativa $H_1 : \boldsymbol{\beta_1} \neq \mathbf{0}$, a estatística de teste *F* segue uma distribuição *F* não central com k_1 e $n-k-1$ graus de liberdade e cujo parâmetro de não centralidade é dado por:

$$\delta = \frac{\boldsymbol{\beta}_1^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}}\left[\mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\right]\mathbf{X}_1\boldsymbol{\beta}_1}{\sigma^2} \quad .$$

3.5. O Teste F de Significância de Restrições Lineares

O teste F pode também ser utilizado para testar a significância de um conjunto de restrições lineares relativas a um modelo de regressão linear múltipla. Nesse caso, a hipótese linear geral a testar é

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \text{ Vs } H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r},$$

onde **R** é uma matriz $m \times (k+1)$ com característica *m*, $m \leq (k+1)$, em que *k* é o número de variáveis regressoras do modelo.

Sob os pressupostos dos erros aleatórios $\varepsilon_i's$ serem idênticos e independentemente distribuídos com $\varepsilon_i \sim N(0,\sigma^2)$, a estatística *F* para testar H_0 assume a seguinte forma:

$$F = \frac{(\mathbf{Rb}-\mathbf{r})^{\mathbf{T}}\left[\mathbf{C}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathbf{T}}\right]^{-1}(\mathbf{Rb}-\mathbf{r})}{ms^2} \sim F_{m,n-k} \quad ,$$

onde $\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$ e $s^2 = \frac{\mathbf{Y}^{\mathbf{T}}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\right)\mathbf{Y}}{n-k-1}$.

Sob a hipótese alternativa $H_1 : \mathbf{R\beta} \neq \mathbf{r}$, a estatística de teste *F* segue uma distribuição *F* não central com *m* e $n-k-1$ graus de liberdade e parâmetro de não centralidade

$$\delta = \frac{(\mathbf{R\beta}-\mathbf{r})^{\mathbf{T}}\left[\mathbf{C}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathbf{T}}\right]^{-1}(\mathbf{R\beta}-\mathbf{r})}{\sigma^2} \ .$$

3.6. O Teste F Usual sobre o Modelo de Regressão Linear em Séries Temporais com Tendência e Sazonalidade

Nesta secção vamos particularizar a aplicação da estatística F no teste à significância de modelos de regressão, em séries temporais com tendência e sazonalidade que, como já vimos no capítulo 2, podem ser escritas como funções lineares no tempo, isto é,

$$Z_t = \beta_1 t + \sum_{i=1}^{p} x_{ti} S_i + \varepsilon_t \, ,$$

com $s_i = S_i - \overline{S}$ e $\beta_0 = \overline{S}$.

Para testar a significância do modelo completo (componentes de tendência e de sazonalidade), definem-se as seguintes hipóteses:

$$H_0 : t = s_1 = s_2 = ... = s_p = 0 \quad H_1 : t \neq 0 \vee \exists i : s_i \neq 0, \quad i = 1,..., p$$

As matrizes a usar para o modelo completo apresentam o seguinte aspeto:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} , \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} , \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ S_1 \\ \vdots \\ S_p \end{bmatrix}_{(p+1)\times 1} , \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

$$\text{com } x_{ti} = \begin{cases} 1 & se \quad t = (j-1)p + i, \text{ para algum } j = 1,..., N_i \\ 0, & \text{caso contrário} \end{cases} .$$

Sob H_0, a estatística *F* seguirá uma distribuição Fisher com $p+1$ e $n-p-1$ graus de liberdade, com

$$F = \frac{\left(\mathbf{Xb} - \overline{\mathbf{Y}}\right)^{\mathbf{T}} \cdot \left(\mathbf{Xb} - \overline{\mathbf{Y}}\right) \Big/ p+1}{\left(\mathbf{Y}^{\mathbf{T}}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\right)\mathbf{Y}\right) \Big/ n-p-1} \quad .$$

A estatística F aplicada sobre este modelo de regressão irá testar a presença de tendência ou de sazonalidade na série temporal em estudo. Caso apenas uma dessas componentes seja significativa no modelo, o teste F, aplicado ao modelo completo, não detetará qual dessas componentes é significativa, pelo que é necessário recorrer-se a um modelo reduzido de regressão linear.

No estudo da significância da componente sazonal, o modelo reduzido de regressão linear a utilizar resulta da eliminação desta componente da equação inicial, ou seja,

$$Z_t = \beta_0 + \beta_1 t_t + \varepsilon_t .$$

O teste de hipóteses a considerar será

$$H_0 : s_1 = s_2 = ... = s_P = 0 \quad \text{Vs} \quad H_1 : \exists i : s_i \neq 0, \quad i = 1,..., p .$$

As matrizes que incorporam este modelo reduzido são as seguintes:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, \quad \mathbf{X_2} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}_{n\times 2}, \qquad \boldsymbol{\beta_2} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

Por fim, a estatística *F*, para testar a significância da componente sazonal em séries temporais com tendência linear, é dada por

$$F = \frac{\left[\mathbf{Y}^{\mathbf{T}}\left(\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{I} - \mathbf{X_2}\left(\mathbf{X_2^T}\mathbf{X_2}\right)^{-1}\mathbf{X_2^T}\right)\mathbf{T}\right] \Big/ (p-1)}{\left[\mathbf{Y}^{\mathbf{T}}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\right)\mathbf{Y}\right] \Big/ (n-p-1)} \sim F_{p-1,n-p-1} \quad .$$

Sob H_0, a estatística *F* segue uma distribuição Fisher com $p-1$ e $n-p-1$ graus de liberdade.

3.7. Erros de Tipo I e Erros de Tipo II Associados ao Teste F

Na aplicação do teste F, tal como acontece nos testes de hipóteses em geral, são consideradas duas hipóteses relativas a parâmetros populacionais – a hipótese nula, representada por H_0, e a hipótese alternativa H_1 que será aceite caso H_0 seja rejeitada. A decisão de aceitação ou rejeição de H_0 é efetuada com base no valor da estatística de teste, que é calculada a partir dos valores de uma amostra retirada da população. Uma vez que estas amostras, à partida, apresentam aleatoriedade, a aceitação ou rejeição de H_0 estará sempre associada a erros e riscos. Estes erros são classificados segundo dois tipos:

Erro tipo I: rejeitar H_0 sendo H_0 verdadeira.
Erro tipo II: Aceitar H_0, sendo H_0 falsa.

Numa tomada de decisão decorrente da aplicação de um teste de hipóteses, podem ocorrer quatro situações distintas, apresentadas no seguinte quadro:

Decisão	Realidade (desconhecida)	
	H_0 é verdadeira	H_0 é falsa
Não rejeitar H0	Decisão correta	**Erro tipo II**
Rejeitar Ho	**Erro tipo I**	Decisão correta

Como a decisão é tomada a partir do valor de uma variável aleatória, a estes erros estarão associadas as probabilidades da sua ocorrência:

$\alpha =$ P(Erro do tipo I) = P(rejeitar H_0 | H_0 é verdadeira), usualmente denominada de nível de significância da amostra.

$\beta =$ P(Erro do tipo II) = P(aceitar H_0 | H_0 é falsa).

O ideal seria conseguir minimizar simultaneamente as probabilidades de erro de tipo I e de tipo II, contudo isto não é possível, uma vez que, quando se reduz a probabilidade de um dos erros, aumenta a probabilidade de ocorrência do outro. Normalmente opta-se por controlar o nível de significância da amostra, ou seja, a probabilidade de ocorrência do erro de tipo I, fixando-se um valor para a probabilidade da sua ocorrência. A partir desta pode ser determinado o valor para a probabilidade do erro de tipo II e a potência de um teste estatístico, que corresponde à probabilidade de rejeitar H_0 quando H_0 é falsa, ou seja,

Potência do teste=P(rejeitar H_0 | H_0 é falsa) = 1- P(aceitar H_0 | H_0 é falsa) $=1-\beta$.

Visto a potência de um teste estatístico representar a probabilidade de se rejeitar corretamente H_0, a comparação das performances de testes estatísticos pode ser feita através da potência de cada um deles. Mantendo o nível de significância constante, será mais eficiente o que apresentar maior potência. É prática comum considerar-se como nível de significância um dos valores 1%, 5% ou 10%.

Em geral, verifica-se ainda que a potência de um teste estatístico diminui à medida que o verdadeiro valor dos parâmetros em causa se aproxima dos valores estabelecidos através de H_0 e que o aumento no tamanho da amostra reduz os valores de α e β.

No caso da aplicação do teste F para testar a significância da componente sazonal de uma série temporal com tendência e sazonalidade, a estatística de teste, sob H_0, segue uma distribuição *F* de Fisher com $p-1$ e $n-p-1$ graus de liberdade. Assim, para um nível de significância de 5%, o chamado valor crítico da distribuição F de Fisher é dado pelo quantil

$$a = F^{-1}{}_{p-1,n-p-1}(0.95). \quad (3.6.2)$$

A rejeição da hipótese inicial irá ocorrer caso o valor da estatística de teste F seja superior ao quantil *a*. A figura seguinte ilustra a função densidade de probabilidade teórica da estatística de teste *F* para 3 e 35 graus de liberdade (numerador e denominador, respetivamente).

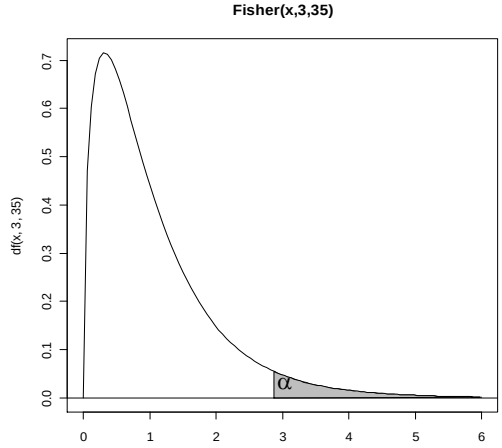


Figura 1. Gráfico da função densidade de probabilidade da distribuição F, para 3 e 35 graus de liberdade do numerador e denominador, respetivamente.

Se o valor da estatística de teste for inferior ao valor crítico, então não se deve rejeitar a H_0. Caso o valor da estatística de teste F se encontre na zona de rejeição, há que considerar a veracidade de H_1, sendo que, nesta situação, a estatística de teste F segue uma distribuição Fisher não central com $p-1$ e $n-p-1$ graus de liberdade e parâmetro de não centralidade δ. A potência do teste F é dada pela probabilidade da estatística de teste F, sob H_1, ser superior ao quantil a (3.6.2) , ou seja,

$$\text{Potência} = P\left(F_{p-1,n-p-1,\delta} > a\right)$$

No gráfico seguinte são apresentadas as funções densidade de probabilidade teóricas da estatística de teste F, sob H_0, com 3 e 35 graus de liberdade e sob H_1, considerando o parâmetro de não centralidade $\delta = 10$.

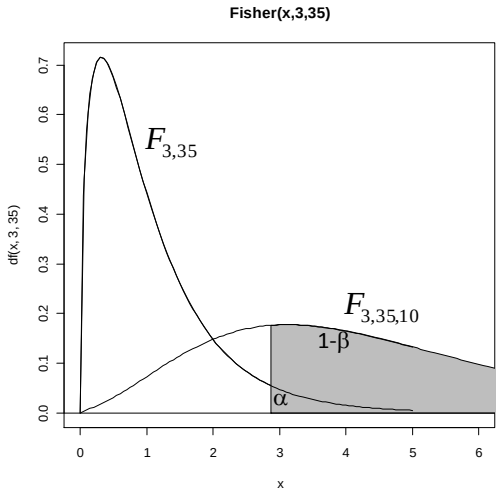


Figura 2. Função densidade de probabilidade da distribuição F não central, para 3 e 35 graus de liberdade do numerador e denominador, respetivamente e parâmetro de não centralidade $\delta = 10$.

Não se deve esquecer que as probabilidades dos erros de tipo I e II, assim como a potência de teste são calculadas partindo da hipótese nula e partindo do princípio que todos os pressupostos de aplicabilidade do teste são verificados, caso contrário os resultados obtidos poderão não ser válidos.

Recorrendo a um estudo de simulação é possível estimar a ocorrência destes dois tipos de erro a um determinado nível de significância nominal, através da aplicação do teste em n séries simuladas sob um conjunto fixo de condições iniciais. O nível de significância empírico é dado pela proporção de rejeições da hipótese nula dado que esta é verdadeira. Se n for elevado é de esperar que, sob H_0, o nível de significância empírico seja um valor próximo do valor de α utilizado. Será tão mais próximo quanto maior for o número de simulações efetuadas. Por outro lado, se as simulações forem

realizadas sob H_1, a proporção de rejeições da hipótese nula corresponderá à potência empírica do teste.

No caso em estudo neste trabalho, para o cálculo do nível de significância empírico, as séries são simuladas sob H_0, ou seja, com os coeficientes sazonais nulos, enquanto para o cálculo da potência empírica, pelo menos um desses coeficientes terá que ser não nulo.

4. Processos Autorregressivos de 1ª Ordem – AR(1)

Como referido no capítulo anterior, a independência dos erros é um dos pressupostos para a aplicação do método dos mínimos quadrados no modelo de regressão linear. No entanto, em muitos casos de séries temporais com tendência e sazonalidade, este pressuposto não é verificado, havendo, portanto, algum tipo de correlação entre os erros do modelo. Com o intuito de modelar esses erros, as séries temporais podem ser vistas como um conjunto de observações de um processo estocástico.

Nesta secção, são abordados os conceitos mais importante para este trabalho no que respeita aos processos estocásticos e, em particular, são apresentadas as características do processo autorregressivo estacionário de 1ª ordem, AR(1).

4.1. Processos Estacionários

Definição 4.1.1. Chama-se *processo estocástico* a qualquer família ou colecção de variáveis aleatórias $\{X(t), t \in T\}$ em que *T* é um conjunto de índices representando o tempo.

Considerando um processo estocástico $\{X(t), t \in R\}$, uma série temporal pode ser vista como um conjunto de observações de um processo estocástico nos instantes $t_1, t_2, ..., t_n$.

O conjunto de índices T, chamado de espaço de parâmetros, nas séries temporais poderá ser um dos seguintes conjuntos: N, Z, R^+ ou R. No caso particular das séries discretas, será considerado *t* inteiro, $t = 0; \pm 1; \pm 2; \ldots$. Por sua vez, a $X(t)$ dá-se o nome de espaço de estados.

Definição 4.1.2. Um processo estocástico $\{X(t), t \in R\}$ diz-se estritamente estacionário sse a distribuição conjunta de $(X(t_1), ..., X(t_n))$ for igual à distribuição conjunta de $(X(t_1 + \delta), ..., X(t_n + \delta))$ para todo o n-úplo $(t_1, ..., t_n)$ e todo o δ, ou seja,

$$F_{(X(t_1),\ldots,X(t_n))}(x_1, ..., x_n) = F_{(X(t_1+\delta),\ldots,X(t_n+\delta))}(x_1, ..., x_n)$$

em todos os pontos $(x_1,...,x_n)$ e em que $F_{(X(t_1),...,X(t_n))}$ representa a função distribuição conjunta de $(X(t_1),...,X(t_n))$.

Pode-se então dizer que, num processo estritamente estacionário, a distribuição de probabilidade de um conjunto qualquer de margens mantém-se inalterada para translações destas no tempo.

Definição 4.1.3. Dado um processo estocástico $X(t)$, tal que $E\left(X(t)^2\right)<+\infty$, define-se:

i) Função valor médio, $\mu(t)$:

$$\mu(t)=E(X(t)).$$

ii) Função de variância, $\sigma^2(t)$:

$$\sigma^2(t)=Var(X(t)).$$

iii) Função de covariância, $\gamma(t_1,t_2)$:

$$\gamma(t_1,t_2)=\text{cov}(X(t_1),X(t_2))=E[X(t_1)\cdot X(t_2)]-\mu(t_1)-\mu(t_2).$$

iv) Função de correlação, $\rho(t_1,t_2)$:

$$\rho(t_1,t_2)=\frac{\gamma(t_1,t_2)}{\sigma(t_1)+\sigma(t_2)}=\frac{\text{cov}(X(t_1),X(t_2))}{\sqrt{\text{var}(X(t_1),X(t_2))}}.$$

Definição 4.1.4. Um processo estocástico diz-se estacionário de 2ª ordem ou estacionário para a covariância sse for tal que, para todo o t, $E\left(X(t)^2\right)<+\infty$,e

i) $E(X(t))=\mu$

ii) $Var(X(t))=\sigma^2$

iii) $Cov(X(t_1),X(t_2))=\gamma(t_1,t_2)=\gamma(|t_2-t_1|)$.

A partir de iii) verifica-se que, num processo estacionário de 2ª ordem, a covariância entre duas margens do processo t_1 e t_2 depende apenas da sua distância temporal. Assim sendo, pode-se dizer que a função de covariância é invariável no tempo, logo pode ser definida através de uma única variável *k*. Esta propriedade também se verifica com a função de correlação.

Funções de autocovariância e de autocorrelação

Definição 4.1.5. Num processo estacionário de 2ª ordem, chama-se:

Função de autocovariância à função:

$$\gamma_k = \mathrm{cov}\big(X(t), X\big(t+k\big)\big).$$

Função de autocorrelação (FAC) à função:

$$\rho_k = \rho\big(X(t), X\big(t+k\big)\big) = \frac{\mathrm{cov}\big(X(t), X\big(t+k\big)\big)}{\sqrt{Var(X(t)) \cdot Var\big(X\big(t+k\big)\big)}} = \frac{\gamma_k}{\gamma_0} \ .$$

A função de autocorrelação tem uma grande relevância na identificação do processo subjacente à componente dos erros de uma série temporal. Nos casos dos dados serem provenientes de amostras, a identificação desse processo terá que ser efetuada a partir das estimativas dos valores da referida função. É natural considerar-se os seguintes estimadores para as funções de autocovariância e autocorrelação, respetivamente:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X}),$$

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}, \quad (4.1.6)$$

em que $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$

O estimador $\hat{\gamma}_k$ é enviesado à esquerda, o mesmo acontecendo com $\hat{\rho}_k$. Quanto maior for *k*, maior será o desvio de $E(\hat{\gamma}_k)$ relativamente a γ_k e de $E(\hat{\rho}_k)$ relativamente a ρ_k, por isso é procedimento comum estimar-se γ_k e ρ_k para os primeiros $\frac{n}{4}$ valores de *k*.

Funções de autocovariância e de autocorrelação parciais

A função de autocorrelação parcial complementa a função de autocorrelação no que respeita à identificação do modelo a que corresponde a componente aleatória da série temporal em estudo. Enquanto a função de autocorrelação (FAC) é uma medida de

associação linear entre X_t e X_{t+k}, independentemente da relação com as variáveis intermédias no processo, a função de autocorrelação parcial (FACP) mede a relação existente entre as variáveis X_t e X_{t+k} depois de retirado o efeito das variáveis intermédias, $X_{t+1}, X_{t+2}, ..., X_{t+k-1}$. A função de autocorrelação parcial pode ser representada por P_K.

Representando por ϕ_{ki} a correlação entre a variável X_{t+k} e X_{t+k-i}, com $i = 1, ..., k$, podemos escrever que:

$$X_{t+k} = \phi_{k1} X_{t+k-1} + \phi_{k2} X_{t+k-2} + ... + \phi_{kk} X_t + e_{t+k},$$

em que o resíduo e_{t+k} é não correlacionado com Z_{t+k-i}, com $i = 1, ..., k$.

A função de autocorrelação parcial é, portanto, a função que, a cada intervalo *k* associa $P_K = \phi_{kk}$.

A estimação da função de autocorrelação parcial, no caso dos processos estacionários de 2ª ordem, poderá ser efetuada a partir das seguintes equações:

$$\rho_j = \phi_{k1} \rho_{j-1} + \phi_{k2} \rho_{j-2} + ... + \phi_{kk} \rho_{j-k},\ j = 1, 2, ..., k,$$

em que os valores de ρ_j serão substituídos pelos de $\hat{\rho}_j$, estimados a partir da amostra. Tem-se então que:

$$\hat{\rho}_j = \hat{\phi}_{k1} \hat{\rho}_{j-1} + \hat{\phi}_{k2} \hat{\rho}_{j-2} + ... + \hat{\phi}_{kk} \hat{\rho}_{j-k},\ j = 1, 2, ..., k.$$

Segundo Dublin (1960), as soluções do sistema de equações anterior podem ser determinadas de forma recursiva, a partir das seguintes expressões:

$$\hat{\phi}_{(k+1)(k+1)} = \frac{\hat{\rho}_{k+1} - \sum_{j=1}^{k} \hat{\phi}_{kj} \hat{\rho}_{k+1-j}}{1 - \sum_{j=1}^{k} \hat{\phi}_{kj} \hat{\rho}_j}$$

e $$\hat{\phi}_{(k+1)j} = \hat{\phi}_{kj} - \hat{\phi}_{(k+1)(k+1)} \hat{\phi}_{k(k+j-1)},\quad j = 1, ..., k$$

4.2. Processo de Ruído Branco e Processo Autorregressivo de 1ª Ordem - AR(1)

Um dos modelos que normalmente está presente na componente dos erros de séries temporais é o chamado processo autorregressivo de 1ª ordem – AR(1), definido a partir de um processo de ruído branco.

Definição 4.2.1. Um processo estocástico $\{a_t, t \in Z\}$ diz-se puramente aleatório ou de ruído branco sse

i) $E(a_t) = \mu$

ii) $Var(a_t) = \sigma_a^2$ (variância constante)

iii) $Cov(a_t, a_{t+k}) = \gamma_k = 0, \quad k = \pm 1, \pm 2, ...$

A partir da definição anterior, pode concluir-se que

$$\rho_k = \begin{cases} 0 & k = \pm 1, \pm 2, ... \\ 1 & k = 0 \end{cases}.$$

Num processo de ruído brando, a sucessão das variáveis aleatórias apresentam correlação nula. É um processo que por si só não tem grande utilidade, todavia é a partir deste que muitos dos processos temporais são definidos. Conclui-se ainda que um processo de ruído branco é um processo estacionário de 2ª ordem.

Definição 4.2.2. Um processo autorregressivo de 1ª ordem AR(1) é um processo definido a partir de um processo puramente aleatório $\{a_t\}$ e que verifica a equação

$$Z_t = \phi Z_{t-1} + a_t,$$

em que a_t é um processo de ruído branco.

Escrevendo a série Z_t a partir do seu primeiro valor e considerando $Z_0 = 0$, tem-se:

$$\begin{aligned} Z_1 &= a_1, \\ Z_2 &= \phi a_1 + a_2, \\ Z_3 &= \phi^2 a_1 + \phi a_2 + a_3, \\ &\vdots \\ Z_t &= \phi^{t-1} a_1 + \phi^{t-2} a_2 + ... + \phi a_{t-1} + a_t. \end{aligned}$$

Assim, nos casos em que $Z_0 = 0$, podemos redefinir Z_t de forma não recursiva, através da seguinte expressão:

$$Z_t = \phi^{t-1}a_1 + \phi^{t-2}a_2 + \ldots + \phi a_{t-1} + a_t = \sum_{j=0}^{t-1} \phi^j a_{t-j} .$$

Dada a correlação existente entre as variáveis deste processo, quando $Z_0 \neq 0$, os valores da série serão afetados de alguma forma por este valor inicial. Contudo, para valores de *t* suficientemente grandes, Z_0 perderá a sua influência no comportamento do processo. Assim sendo, considerando que o início da série é dado por Z_{-N} , com *N* a tender para o infinito, tem-se que

$$Z_t = \sum_{j=-\infty}^{t} \phi^{t-j} a_j = \sum_{j=0}^{+\infty} \phi^j a_{t-j} .$$

Contudo, para Z_t poder ser representado pelas séries representadas na expressão anterior, é preciso que essas séries sejam convergentes em média quadrática, o que acontece se $|\phi| < 1$. Deduz-se ainda que o processo AR(1) é estacionário sse $|\phi| < 1$.

Do resultado anterior pode concluir-se que, num processo AR(1), se $|\phi| < 1$, a covariância e a correlação entre dois pontos do processo, t_1 e t_2, dependem apenas da distância temporal, ou seja:

$$\gamma(t_1, t_2) = \gamma(|t_2 - t_1|) ,$$

$$\rho(t_1, t_2) = \rho(|t_2 - t_1|) .$$

Propriedade 4.2.3. Num processo AR(1), tem-se $E(Z_t) = 0$ e $\text{var}(Z_t) = \dfrac{\sigma_a^2}{1-\phi^2}$.

Demonstração:

Como $E(a_t) = 0$, para $t \in T$, $E(Z_t) = E\left(\sum_{j=0}^{+\infty} \phi^j a_{t-j}\right) = \sum_{j=0}^{+\infty} E\left(\phi^j a_{t-j}\right) = \sum_{j=0}^{+\infty} \phi^j E\left(a_{t-j}\right) = 0$

$$\text{var}(Z_t) = \text{var}\left(\sum_{j=0}^{+\infty} \phi^j a_{t-j}\right) = \sum_{j=0}^{+\infty} \text{var}\left(\phi^j a_{t-j}\right) = \sum_{j=0}^{+\infty} \phi^{2j}\, \text{var}\left(a_{t-j}\right) = \sigma_a^2 \sum_{j=0}^{+\infty} \phi^{2j}$$

Para $|\phi| < 1$, a série geométrica $\sum_{j=0}^{+\infty} \phi^{2j}$ é convergente, sendo a sua soma $\dfrac{1}{1-\phi^2}$,

logo, $\text{var}(Z_t) = \dfrac{\sigma_a^2}{1-\phi^2}$ ■

Em (4.1.5) foram apresentadas as funções de autocovariância e de autocorrelação para qualquer processo estacionário de 2ª ordem, todavia existem expressões mais simplificadas para o caso do processo AR(1).

A função de autocovariância de um processo AR(1) é dada pela expressão

$$\gamma_k = \phi^{|k|} \cdot \frac{\sigma_a^2}{1-\phi^2} \text{ com } k = 0,\pm1,\pm2,\ldots .$$

Demonstração:

Multiplicando ambos os membros da equação $Z_t = \phi Z_{t-1} + a_t$ por $Z_{t-k}, k > 0$ obtém-se,

$$Z_t \cdot Z_{t-k} = \phi Z_{t-1} \cdot Z_{t-k} + a_t \cdot Z_{t-k} .$$

Tomando valores médios,

$$E(Z_t \cdot Z_{t-k}) = \phi E(Z_{t-1} \cdot Z_{t-k}) + E(a_t \cdot Z_{t-k}).$$

Como as variáveis a_t e Z_{t-k} são independentes com valor médio nulo,

$$E(a_t \cdot Z_{t-k}) = E(a_t)E(Z_{t-k}) = 0 .$$

Assim, $\gamma_k = \text{cov}(Z_t, Z_{t-k}) = E[Z_t, Z_{t-k}] - \mu(t) - \mu(t-k)$

$$\gamma_k = \phi E(Z_{t-1}, Z_{t-k})$$

$$\gamma_k = \phi^2 E(Z_{t-2}, Z_{t-k})$$

$$\gamma_k = \phi^k E(Z_{t-k}, Z_{t-k}) = \phi^k \text{ var}(Z_{t-k}) = \phi^k \gamma_0 .$$

Como $\gamma_0 = \text{var}(Z_t) = \frac{\sigma_a^2}{1-\phi^2}$,

$$\gamma_k = \phi^k \cdot \frac{\sigma_a^2}{1-\phi^2}$$

Como $\gamma^k = \gamma^{-k}$, então $\gamma_k = \phi^{|k|} \cdot \frac{\sigma_a^2}{1-\phi^2}$ com $k = 0,\pm1,\pm2,\ldots .$ ■

Por sua vez, a função de autocorrelação de um processo AR(1) é dada por:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\phi^{|k|} \cdot \gamma_0}{\gamma_0} = \phi^{|k|}, \quad k = 0,\pm1,\pm2,\ldots .$$

Neste processo, cada variável aleatória Z_t está correlacionada com a variável aleatória definida no instante anterior, Z_{t-1}. Uma vez que o processo AR(1) é

estacionário de 2ª ordem, ou seja, estacionário para a variância, o valor do coeficiente de correlação entre duas variáveis consecutivas no tempo, Z_t e Z_{t-1},dado por ρ_1, é constante. Este valor corresponde ao valor da constante ϕ definida na equação deste modelo.

Naturalmente que um estimador para o parâmetro autorregressivo ϕ será

$$\hat{\phi} = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \hat{\rho}_1 . \quad (4.2.4)$$

O estimador $\hat{\phi}$ em (4.2.4) engloba-se nos denominados estimadores de *Yule-Walker*. Nos capítulos 6 e 7, estudo de simulação e análise de séries reais, respetivamente, é utilizado este estimador para determinar a estimativa do respetivo parâmetro.

No que concerne ao modelo de regressão linear quando os erros apresentam autocorrelação serial, é importante salientar que o estimador dos mínimos quadrados ordinários continua a ser não enviesado e consistente (a estimativa converge para o valor do parâmetro) no entanto perde eficiência, deixando de ter variância mínima. Este aumento da variância do estimador, face ao do modelo sem correlação, leva também a que os testes estatísticos de significância dos parâmetros e os limites de confiança dos mesmos deixem de ser os corretos. Esta situação é mais problemática para valores de correlação maiores.

5. Estatística de Teste F em Séries com Autocorrelação

5.1. Estimação pelos Mínimos Quadrados Generalizados

A independência dos erros é um dos pressupostos na aplicação do teste *F* usual para testar a significância de um conjunto de parâmetros num modelo linear. Em geral, se as observações forem correlacionadas no tempo, a estatística *F* usual não segue uma distribuição *F*, pelo que as conclusões resultantes da sua aplicação poderão deixar de ser válidas.

Quando os erros apresentam autocorrelação, os coeficientes de regressão podem ser estimados através do método dos mínimos quadrados generalizados (MQG). Nesta situação, sob H_0, a estatística *F* para testar a significância do modelo tem distribuição conhecida e será apresentada neste capítulo.

Consideremos o seguinte modelo de regressão linear:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \ (5.1.1)$$

em que, $\mathbf{Y}$ representa o vetor da variável dependente,

$\mathbf{X}$ representa a matriz das variáveis independentes (sem dependência linear),

$\boldsymbol{\beta}$ representa o vetor dos coeficientes de regressão,

$\boldsymbol{\varepsilon}$ representa o vetor dos erros aleatórios, assumido como normalmente distribuído com média zero e matriz de covariância $\boldsymbol{\Omega}$ simétrica e definida positiva, $\boldsymbol{\Omega} = \mathbf{C}\mathbf{C}^{\mathbf{T}}$, com $\mathbf{C}$ triangular inferior. Exige-se ainda que $\boldsymbol{\varepsilon}$ e $\mathbf{X}$ sejam independentes.

Quando os erros são gerados por um processo autorregressivo de 1ª ordem, $\varepsilon_t = \phi\varepsilon_{t-1} + a_t$, com $a_t = N\left(0, {\sigma_a}^2\right)$, a matriz $\boldsymbol{\Omega}$ é dada por $\Omega = \dfrac{{\sigma_a}^2}{1-\phi^2} W$, onde

$$W = \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{bmatrix} .$$

Sob H_0, o fator de escala em $\boldsymbol{\Omega}$, ${\sigma_a}^2 / 1-\phi^2$, pode ser ignorado, passando esta matriz a ser escrita unicamente a partir do parâmetro autorregressivo ϕ. A estimativa

$\hat{\boldsymbol{\Omega}}$ da matriz de covariância é obtida substituindo na matriz anterior o valor do parâmetro autorregressivo pela respetiva estimativa $\hat{\phi}$.

Com a matriz $\boldsymbol{\Omega}$ conhecida, a aplicação do método dos mínimos quadrados ordinários ao sistema (5.1.1) depois da sua pré-multiplicação por C^{-1}, resulta no estimador de β pelo método dos mínimos quadrados generalizados, que é dado por

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{\Omega}^{-1}\mathbf{Y}.$$

Este estimador é o melhor, linear e não enviesado e é idêntico ao estimador dos mínimo quadrados ordinários sse $\Omega = \mathbf{X}\Gamma\mathbf{X}^{\mathbf{T}} + \mathbf{Z}\theta\mathbf{Z}^{\mathbf{T}} + \sigma^2\mathbf{I}$, onde Γ, θ, e σ^2 são arbitrários e $\mathbf{Z}$ é uma matriz tal que $\mathbf{X}^{\mathbf{T}}\mathbf{Z} = \mathbf{0}$ (Rao 1967).

Em particular, quando os erros aleatórios são gerados por um processo AR(1), o estimador de $\boldsymbol{\beta}$ pelo método dos mínimos quadrados generalizados é dado por:

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{-1}\mathbf{Y}. \quad (5.1.2)$$

Se a matriz **W** não for conhecida, mas se for plausível os erros serem modelados por um processo autorregressivo de 1ª ordem, há que testar se, de facto, a autocorrelação nos resíduos é significativamente superior a 0, o que pode ser efetuado pelo teste de Durbin-Watson ou por um teste de Portmanteau.

De seguida é apresentado um processo de estimação a duas etapas para a estimação dos coeficientes de regressão, quando ϕ é desconhecido.

Estimação dos coeficientes de regressão por um processo a duas etapas.

A partir da substituição de $\hat{\phi}$ na matriz $\mathbf{W}^{-1}$, é calculado o estimador $\hat{\mathbf{W}}^{-1}$ e seguidamente

$$\hat{\mathbf{b}} = \left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}^{-1}\mathbf{Y}. \quad (5.1.3)$$

Este procedimento poderá ter que ser repetido como um processo iterativo até se atingir uma relativa estabilidade dos estimadores $\hat{b}$ e $\hat{\phi}$.

A iteração começa com uma estimação pelos mínimos quadrados ordinários de $\mathbf{b_0} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$. São então calculados $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b_0}$, $\hat{\phi}$ e $\hat{\mathbf{b}}$ (a partir de 5.1.3).

Calcula-se novamente $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}$, $\hat{\phi}$ (a partir deste último e) e $\hat{\mathbf{b}}$. Este processo repetir-se-á até que as variações em $\hat{\phi}$ e $\hat{\mathbf{b}}$ sejam menores que um dado valor estabelecido.

5.2. Método de Estimação de Cochrane-Orcutt

Alternativamente ao método dos MQG, pode-se recorrer a uma transformação nos dados para contornar o problema da presença da autocorrelação. O modelo com as variáveis transformadas apresenta homocedasticidade dos resíduos e pode ser estimado pelos MQO. Este procedimento é conhecido pelo método de Cochrane-Orcutt, cuja descrição da transformação de variáveis é apresentada de seguida:

O modelo de regressão linear estabelece que:

$$y_t = \beta_0 + \beta_1 x_{1(t)} + ... + \beta_k x_{k(t)} + \varepsilon_t \qquad (5.2.1)$$

e $$y_{t-1} = \beta_0 + \beta_1 x_{1(t-1)} + ... + \beta_k x_{k(t-1)} + \varepsilon_{t-1}, \quad (5.2.2)$$

onde, ε_t é um processo AR(1): $\varepsilon_t = \phi\varepsilon_{t-1} + a_t$, com a_t processo de ruído branco.

Multiplicando ambos os membros de (5.2.2) por ϕ, tem-se

$$\phi y_{t-1} = \phi\beta_0 + \phi\beta_1 x_{1(t-1)} + ... + \phi\beta_k x_{k(t-1)} + \phi\varepsilon_{t-1}. \quad (5.2.3)$$

Subtraindo (5.2.1) por (5.2.3), obtém-se

$$\left(y_t - \phi y_{t-1}\right) = \beta_0(1-\phi) + \beta_1\left(x_{1(t)} - \phi x_{1(t-1)}\right) + ... + \beta_k\left(x_{k(t)} - \phi x_{k(t-1)}\right) + \left(\varepsilon_t - \phi\varepsilon_{t-1}\right).$$

Como $\varepsilon_t - \phi\varepsilon_{t-1} = a_t$, então

$$\left(y_t - \phi y_{t-1}\right) = \beta_0(1-\phi) + \beta_1\left(x_{1(t)} - \phi x_{1(t-1)}\right) + ... + \beta_k\left(x_{k(t)} - \phi x_{k(t-1)}\right) + a_t,$$

o que pode ser expresso por

$$\Delta y_t = \beta_0(1-\phi) + \beta_1\Delta x_{1t} + \cdots + \beta_k\Delta x_{kt} + a_t,$$

com $\Delta y_t = y_t - \phi y_{t-1}$ e $\Delta x_{i(t)} = x_{i(t)} - \phi x_{i(t-1)}$.

Nota: à exceção de β_0 , os restantes parâmetros β_i permanecem inalterados.

O procedimento de Cochrane-Orcutt é iterativo e consta dos seguintes passos:

1) Estimar o modelo de regressão pelos MQO e a partir deste calcular os resíduos e_t ;
2) A partir dos resíduos e_t obter a estimativa $\hat{\phi}$ do parâmetro autorregressivo;
3) Através da estimativa $\hat{\phi}$, calcular $\Delta y_t = y_t - \hat{\phi} y_{t-1}$ e $\Delta x_{i(t)} = x_{i(t)} - \hat{\phi} x_{i(t-1)}$.
4) Proceder à regressão linear de Δy_t em $\Delta x_{i(t)}$. A constante nesta regressão será $\beta_0\left(1-\hat{\phi}\right)$. Calcular os novos resíduos a partir desta regressão.
5) Usar os novos resíduos para obter uma nova estimativa $\hat{\phi}$ do parâmetro autorregressivo;
6) Repetir os passos a partir de 3) até se obter convergência.

Com este procedimento os valores da primeira observação são perdidos, o que, numa amostra reduzida, poderá ser desvantajoso.

5.3. Teste F com Estimação pelo Método dos Mínimos Quadrados Generalizados

Quando a estimação dos parâmetros do modelo é efetuada pelos MQG, a expressão da estatística *F* para testar um conjunto de *m* restrições lineares em $\boldsymbol{\beta}$, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, em que $\mathbf{R}$ é uma matriz de dimensão $m \times k$, com característica $m \le k$ e $\mathbf{r}$ um vetor de dimensão $m \times 1$, é dada por:

$$F = \frac{(\mathbf{KY} - \mathbf{r})^{\mathbf{T}} \mathbf{Q} (\mathbf{KY} - \mathbf{r})}{\mathbf{Y}^{\mathbf{T}} \mathbf{M} \mathbf{Y}} \quad , \quad (5.3.1)$$

onde $\mathbf{K} = \mathbf{R}\left(\mathbf{X}^{\mathbf{T}} \boldsymbol{\Omega}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathbf{T}} \boldsymbol{\Omega}^{-1}$,

$$\mathbf{Q} = \left[\mathbf{R}\left(\mathbf{X}^{\mathbf{T}} \boldsymbol{\Omega}^{-1} \mathbf{X}\right)^{-1} \mathbf{R}^{\mathbf{T}} \right]^{-1} \Big/ m \quad ,$$

$$\mathbf{M} = \left[\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \mathbf{X}\left(\mathbf{X}^{\mathbf{T}} \boldsymbol{\Omega}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathbf{T}} \boldsymbol{\Omega}^{-1} \right] \Big/ (T-k) \quad .$$

Sob a hipótese nula $H_0 : \mathbf{R\beta} = \mathbf{r}$, a estatística de teste F em (5.3.1) segue uma distribuição $F_{m,T-k}$. Para além disso, sob a hipótese alternativa, $H_1 : \mathbf{R\beta} \neq \mathbf{r}$, a expressão (5.3.1) segue uma distribuição *F* não central com *m* e *T-k* graus de liberdade e parâmetro de não centralidade $\upsilon = \mathbf{p}^{\mathbf{T}}\mathbf{Q}\mathbf{p}$, com $\mathbf{p} = \mathbf{(R\beta - r)}$.

Quando $\mathbf{\Omega} = \mathbf{I}$, ou seja, quando os erros aleatórios forem independentes, a expressão da estatística *F* em (5.3.1) traduz-se na estatística *F* usual, aplicada segundo os mínimos quadrados ordinários, para testar restrições lineares relativas aos coeficientes.

Quando é utilizada uma estimativa $\hat{\Omega} \neq \Omega$ para a matriz de covariância, a estatística *F* assume a seguinte expressão:

$$\hat{F} = \frac{(\mathbf{p} + \mathbf{K}_1\boldsymbol{\eta})^{\mathbf{T}}\mathbf{Q}_1(\mathbf{p} + \mathbf{K}_1\boldsymbol{\eta})}{\boldsymbol{\eta}^{\mathbf{T}}\mathbf{M}_1\boldsymbol{\eta}} \quad (5.3.2)\,,$$

onde $\boldsymbol{\eta} = \mathbf{C}^{-1}\boldsymbol{\mu} \sim N(0, I)$,

$\mathbf{p} = \mathbf{R\beta} - \mathbf{r}$,

$\mathbf{K}_1 = \mathbf{R}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{\Omega}}^{-1}\mathbf{C}$,

$\mathbf{Q}_1 = \left[\mathbf{R}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathbf{T}}\right]^{-1} \Big/ m$,

$\mathbf{M} = \left[\mathbf{\Omega}^{-1} - \hat{\mathbf{\Omega}}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{\Omega}}^{-1}\right]\mathbf{C} \Big/ (T-k)$.

Em geral, quando $\hat{\mathbf{\Omega}} \neq \mathbf{\Omega}$, a expressão (5.3.2) não segue uma distribuição *F*, pelo que os valores das probabilidades $P\left(\hat{F} \leq q\right)$ terão que ser determinados por outro processo que não o recurso direto à função distribuição de probabilidade. Existem vários métodos para calcular estas probabilidades, sob H_0, como por exemplo o procedimento Imhof's (1961) ou o procedimento apresentado em Palm e Sneek (1981). Dado estes procedimentos não fazerem parte da temática principal que nos propusemos desenvolver, o seu estudo não foi incluído neste trabalho.

De referir que, quando a matriz $\hat{\mathbf{\Omega}}$ é não estocástica, sob H_0, o fator de escala σ^2 em $\hat{\mathbf{\Omega}}$ não tem influencia na distribuição de $\hat{F}$. Portanto, quando em $\hat{\mathbf{\Omega}}$ apenas é estimado o fator de escala σ^2, é preferível, do ponto de vista computacional, parametrizar o processo para ε em (5.1.1) como $N\left(0, \sigma^2\mathbf{\Omega}\right)$, pois, sob H_0, o valor da

estatística $\hat{F}$ não depende da estimação de σ^2. Todavia, sob H_1, uma sobrestimação de σ^2 resulta numa subestimação da função potência e vice-versa, com o parâmetro de não centralidade a variar proporcionalmente a σ^{-2}.

5.4. Comportamento do Teste F quando os Erros são Autocorrelacionados de Ordem 1

Nesta secção são apresentados alguns resultados de estudos relativos ao comportamento da estatística F no modelo de regressão linear quando os erros são autocorrelacionados de 1ª ordem.

Kiviet (1979) obteve intervalos exatos para o valor crítico da estatística *F* usual, quando os coeficientes de regressão são estimados pelos MQO, em modelos lineares em que os erros são gerados por processos ARMA simples. Ficou claro nos seus resultados que as conclusões da aplicação do teste F são muito afetadas quando a autocorrelação nos erros não é considerada no cálculo da estatística de teste.

Palm e Sneek (1984) estenderam a análise da distribuição exata da estatística de teste F para situações em que a autocorrelação dos erros é tomada em linha de conta através da estimação pelos MQG. Nos seus estudos, estes autores analisaram a influência que uma incorreta estimação do parâmetro autorregressivo teria no valor da probabilidade de erro de tipo I. Dos resultados obtidos, ficou patente que a distribuição exata da estatística de teste por vezes difere substancialmente da incorretamente assumida distribuição F. A real distribuição da estatística *F* move-se para a esquerda quando a autocorrelação é sobrestimada ($\hat{\rho} > \rho$) e desloca-se para a direita quando esta é subestimada ($\hat{\rho} < \rho$). Em particular, a verdadeira probabilidade de ocorrência de erro do tipo I cresce substancialmente quando $\rho = 0.9$ é severamente subestimada.

Palm e Sneek concluíram também que a verdadeira probabilidade de ocorrência de erro de tipo I muda com os coeficientes de regressão que estão a ser testados e aparentam piorar com o aumento do número de restrições. Esta probabilidade depende ainda da matriz dos regressores e das restrições a serem testadas.

Segundo Palm e Sneek, a estimação pelos mínimos quadrados generalizados nem sempre é a melhor solução na presença de autocorrelação nos resíduos, pelo que deve haver muito cuidado quando são usados testes de significância em regressões cujos resíduos do modelo são altamente correlacionados. Para uma inferência mais correta,

em conjunto com o teste F, deverá ser aplicado um teste de independência dos resíduos como, por exemplo, o teste de Durbin-Watson.

6. Estudos de Simulação

A simulação de Monte Carlo é uma técnica muito utilizada para avaliar o comportamento de testes estatísticos, especialmente no que respeita a estimar erros de tipo I e II e função potência do teste. A técnica é bastante útil quando não existe uma forma explícita de se calcular a função potência do teste a partir de hipóteses alternativas bem definidas e no caso dos testes paramétricos, quando a função referida passa a ser desconhecida e difícil de obter se forem alteradas algumas das condições iniciais. Neste capítulo apresentamos um estudo de simulação deste tipo, visando avaliar o comportamento do teste F na deteção de sazonalidade em modelos lineares, quando não são verificados alguns dos pressupostos da sua aplicabilidade e de garantia das suas propriedades estatísticas, nomeadamente, quando falha a normalidade dos erros e/ou independência entre as observações. Para a componente estocástica (componente do erro) é considerada sempre a estrutura de processo autorregressivo de 1ª ordem, AR(1), definido a partir da expressão

$$\varepsilon_t = \phi\varepsilon_{t-1} + a_t \text{, com } |\phi| < 1,$$

em que a_t é um processo de ruído branco e $t = 1,...,n$.

O processo de ruído branco a_t é gerado sob duas distribuições: a normal e a gama (de média nula). A escolha recaiu sobre estas distribuições pois, se, por um lado, o teste F apresenta propriedades ótimas para os erros normais, por outro lado, considerar a distribuição gama vai permitir avaliar o comportamento do teste F para vários níveis de assimetria (à direita).

A metodologia adotada para o processo de simulação de cada série começa pela simulação aleatória de um processo de ruído branco com uma das duas distribuições mencionadas. A partir deste, constrói-se um processo estacionário ε_t, do tipo AR(1), com um determinado nível de correlação entre valores consecutivos. Posteriormente, são adicionadas as componentes de tendência e de sazonalidade ao modelo, obtendo-se, então, um conjunto de valores ordenados no tempo que respeitam o seguinte modelo:

$$Z_t = \beta_1 t + \sum_{i=1}^{P} x_{ti} S_i + \varepsilon_t ,$$

A dimensão da série simulada e os valores dos parâmetros das distribuições consideradas para a geração das séries dos resíduos variaram consoante as necessidades verificadas para cada caso. Para o parâmetro autorregressivo ϕ utilizou-se os valores: $\phi = 0$, $\phi = 0.25$, $\phi = 0.5$, $\phi = 0.75$ e $\phi = 0.9$. Sabemos por resultados de estudos anteriores que a existência de autocorrelação deve ser considerada significativa para valores ≥ 0.5, no entanto consideraram-se valores inferiores na simulação que possibilitam analisar com mais detalhe a variação nos valores do nível de significância e potência do teste. Não foram consideradas correlações negativas no estudo de simulação porque é pouco comum encontrar-se, em dados de variáveis reais, processos autorregressivos de ordem 1 de memória negativa. Na aplicação do teste F, para testar a significância da componente sazonal, foi considerado um nível de significância nominal de 5%.

Numa primeira instância, para representar os coeficientes sazonais, utilizou-se uma sequência de valores com período 4, s_i=(-26, 3, 20, 3). Estes valores são aproximações dos coeficientes estimados pelo método dos MQO, numa regressão linear de uma série de valores reais referentes à produção trimestral de tijolos, durante o período compreendido entre Janeiro de 1956 e Dezembro de 1973 [1]. Para o declive obteve-se o valor aproximado $\beta_1 = 4$ e, como intersecção com o eixo vertical, $\beta_0 = 177$.

A análise dos resultados obtidos com as simulações baseia-se no seguinte:

- <u>Estudo comparativo entre as estimativas de erro de tipo I associadas à aplicação do teste F:</u> num teste estatístico é usual fixar-se o nível de significância, ou seja, a proporção de rejeições incorretas da hipótese nula, valor este que aparece associado à decisão decorrente da aplicação do teste. Sendo o nível de significância empírico a proporção de rejeições de H_0 (H_0:s_1=s_2=…=s_p=0), quando a série gerada não tem sazonalidade, a sua comparação com o nível de significância nominal é importante no estudo da *performance* do teste *F* na detecção de sazonalidade no modelo linear. Uma diferença significativa entre estes dois valores quererá dizer que a real distribuição da estatística de teste difere significativamente da distribuição considerada sob H_0. Para além disso, dada a relação que existe entre o erro de tipo I e o erro de tipo II, a alteração de um destes irá afetar o valor do outro. Assim, um aumento do nível de significância

[1] Dados retirados do sítio <u>http://robjhyndman.com/TSDL/production/</u>

empírico face ao nominal levará possivelmente a uma potência de teste mais elevada ou vice-versa. Portanto, os resultados de potência de um teste estatístico não devem aparecer desligados do nível de significância que lhe está associado. De referir ainda que para o nível de significância nominal considerado ao longo do estudo, 5%, o intervalo de confiança para a o nível de significância empírico é [0,043; 0,059].

- Estudo comparativo da potência do teste *F*. A potência do teste F na deteção de sazonalidade é estimada pela proporção de rejeições da hipótese nula quando é imposta uma componente sazonal ao modelo.

No que diz respeito à deteção de tendência em modelos de regressão linear simples ou múltipla, refira-se um trabalho de Ramos, Rosário (2006), no qual, através de simulação de Monte Carlo, conclui-se que a fraca estimação da autocorrelação é a principal responsável pelo comportamento insuficiente nos testes de detecção de tendência, paramétricos ou não paramétricos. Este é mais grave para os declives mais próximos de zero e quando a autocorrelação toma valores mais elevados, principalmente quando as séries são de pequena dimensão. Nos processos autorregressivos AR(1) a detecção de tendência mostra-se idêntica para modelos com e sem sazonalidade. Quando o parâmetro de autocorrelação é subestimado verifica-se um aumento na probabilidade de rejeitar H_0, quando esta é verdadeira.

De referir também o trabalho realizado por Palm e Sneek (1984), no qual foi investigada a influência que a estimação errada do parâmetro autorregressivo poderá ter na *performance* do teste *F*, no ajustamento a um modelo linear, quando a componente dos erros é gerada por um processo AR(1) com a_t normal. Os resultados mostraram que uma sobrestimação de ϕ tem como consequência a diminuição do nível de significância empírico, enquanto uma subestimação de ϕ provoca o efeito contrário.

6.1. Processo AR(1) em que os a_t têm Distribuição Normal

É de conhecimento geral que a dependência entre as observações altera significativamente as propriedades dos estimadores pelos MQO numa regressão linear, assim como o comportamento do teste F usual na avaliação da qualidade do modelo.

Neste estudo, pretende-se averiguar de que forma a correlação entre as observações afeta a *performance* do teste F na deteção de uma componente sazonal num modelo de regressão linear, quando a_t segue uma distribuição normal com valor médio 0 e variância σ^2.

Para o caso de $\phi = 0$, a série dos erros ε_t é normalmente distribuída apresentando homocedasticidade e independência, o que garante, à partida, os pressupostos de aplicabilidade do teste F, pelo que os resultados decorrentes da análise do estudo de potência e do nível de significância empírico deverão coincidir com os resultados teóricos que se esperam. Nas situações em que o parâmetro autorregressivo for não nulo, a série dos resíduos ε_t, para além apresentar correlação, poderá distanciar-se significativamente de uma distribuição normal, o que pode afetar, de forma significativa, as probabilidades de aceitar e de rejeitar erradamente H_0.

Com um carácter ilustrativo, num primeiro caso, é apresentada uma série temporal com os coeficientes definidos na introdução deste capítulo, em que a série é simulada sem componente de erro.

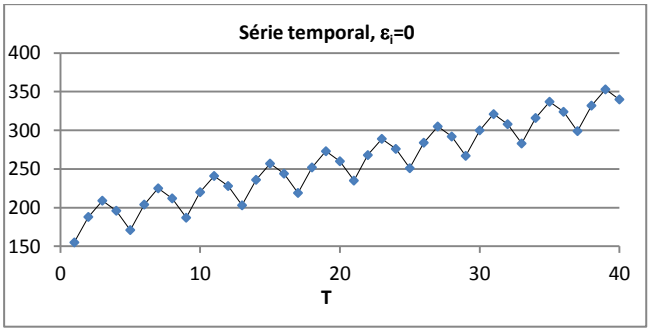


Figura 3. Série temporal com tendência e sazonalidade simulada com $\beta_1 = 4$, $s_i = (-26,3,20,3)$, $\beta_0 = 177$ e $\varepsilon_i = 0$.

A partir da análise de resultados que não foram incluídos neste texto, constatou-se que, para desvios padrão reduzidos ($\sigma < 10$), o teste F detetou a presença de sazonalidade em todas ou praticamente todas as simulações realizadas. Portanto, para $\sigma < 10$, mesmo para valores elevados do parâmetro ϕ, a potência do teste F é aproximadamente 1. Assim sendo, o estudo de potência, para os coeficientes sazonais

utilizados, foi efetuado para valores de σ iguais ou superiores a 10. Na figura 3 pode-se visualizar dois exemplos de séries temporais para $\sigma = 10$, com $\phi = 0$ e $\phi = 0.9$, respetivamente. Nestas, observa-se que as regressões obtidas apresentam um bom ajustamento às séries simuladas, nomeadamente no que respeita à sazonalidade.

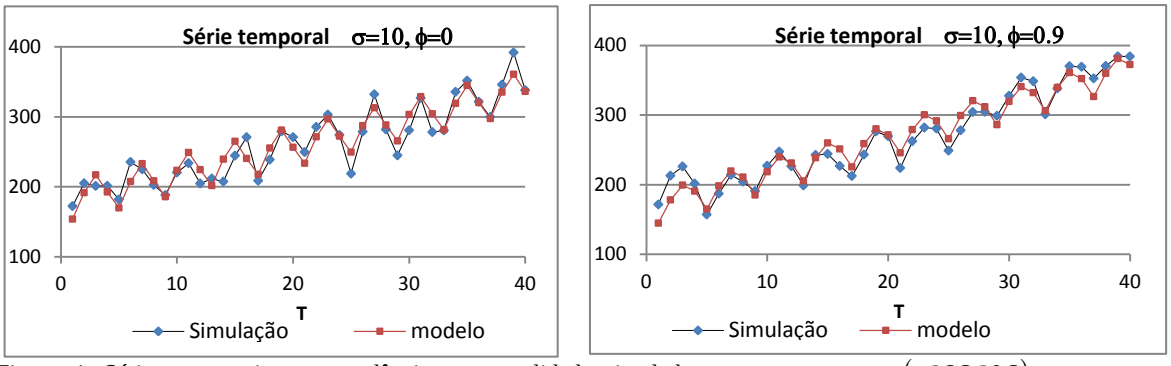


Figura 4. Séries temporais com tendência e sazonalidade simuladas com $\beta_1 = 4$, $s_i = (-26,3,20,3)$, $\beta_0 = 177$ e componente dos erros segundo um processo AR(1) com distribuição Normal(0, $\sigma = 10$). Gráficos para $\phi = 0$ e $\phi = 0.9$.

Por sua vez, da análise à figura 5, verifica-se que, para um desvio padrão da série de ruído branco $\sigma = 20$, a correlação na série dos resíduos afeta de forma mais séria a qualidade do ajustamento do modelo obtido pela regressão linear, em particular da componente sazonal.

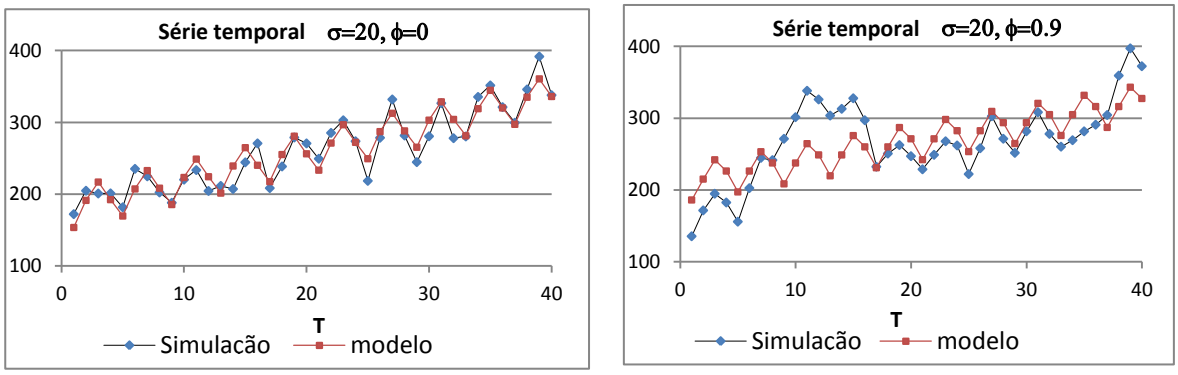


Figura 5. Séries temporais com tendência e sazonalidade simuladas com $\beta_1 = 4$, $s_i = (-26,3,20,3)$, $\beta_0 = 177$ e componente dos erros segundo um processo AR(1) com distribuição Normal(0, $\sigma = 20$). Gráficos para $\phi = 0$ e $\phi = 0.9$.

Nas situações apresentadas constatou-se que, quanto maiores os valores do parâmetro ϕ, pior será o ajuste do modelo à série simulada, sendo esta discrepância mais evidente em séries construídas a partir de ruídos brancos com maior variância. Nestas situações, as discrepâncias entre os valores simulados e os valores obtidos a partir de regressão irão afetar a probabilidade de se rejeitar H_0, dado que esta é falsa, ou seja, irão afetar a probabilidade de se aceitar corretamente a existência de sazonalidade.

Assim, é de prever que um aumento do valor do parâmetro ϕ produza uma redução na potência do teste F.

Para se perceber melhor o comportamento do teste F para variações do desvio padrão da série de ruído branco, do parâmetro de autocorrelação, da dimensão da amostra e do parâmetro de tendência, é realizado um estudo por simulação e são calculados a potência empírica e o nível de significância empírico do teste para diferentes valores dos parâmetros referidos.

6.1.1. Variações do Parâmetro de Tendência e da Constante β_0

À partida, a *performance* da estatística F usual no teste à significância da componente sazonal não deverá ser afetada por flutuações do valor do parâmetro de tendência β_1, desde que os restantes parâmetros permaneçam constantes. Este facto pode ser constatado na tabela 3, onde são apresentados os resultados da potência empírica, assim como o nível de significância empírico resultantes da aplicação do teste F a séries temporais, simuladas para alguns valores de β_1 . Para além dos coeficientes sazonais, fixou-se $\beta_0 = 177$ e T=40 .

	$\beta_1 = 0$		$\beta_1 = 0.5$		$\beta_1 = 1$		$\beta_1 = 20$	
	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível signif.
$\phi = 0$	0,996	0,051	0,992	0,043	0,994	0,048	0,991	0,040
$\phi = 0.25$	0,989	0,024	0,992	0,029	0,991	0,030	0,993	0,033
$\phi = 0.5$	0,980	0,010	0,983	0,008	0,980	0,009	0,984	0,009
$\phi = 0.75$	0,904	0,001	0,908	0,002	0,886	0,001	0,906	0,002
$\phi = 0.9$	0,741	0,001	0,762	0,000	0,744	0,000	0,733	0,002

Tabela 3. Potência e nível de significância empíricos do teste F, sobre a componente sazonal do modelo em que os erros são um processo AR(1) com ruído branco Normal(0, $\sigma = 20$) . T=40. Resultados para $\beta_1 = 0$, $\beta_1 = 0{,}5$, $\beta_1 = 1$ e $\beta_1 = 20$.

Pela análise da tabela 3, podemos verificar que, quer a potência de teste, quer o nível de significância empírico, não sofrem alterações significativas para diferentes valores do parâmetro de tendência β_1. Estudo semelhante foi realizado para se investigar a influência que a variação da constante β_0 poderá ter nos resultados do teste *F*. Assim, para os parâmetros considerados no caso anterior e para $\beta_1 = 4$, obteve-se os resultados apresentados na tabela 4.

	$\beta_0 = -100$		$\beta_0 = 0$		$\beta_0 = 100$		$\beta_0 = 500$	
	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível signif.
$\phi = 0$	0,988	0,060	0,990	0,060	0,993	0,046	0,992	0,046
$\phi = 0.25$	0,993	0,027	0,990	0,030	0,990	0,031	0,987	0,024
$\phi = 0.5$	0,977	0,006	0,977	0,010	0,981	0,009	0,985	0,007
$\phi = 0.75$	0,902	0,002	0,901	0,001	0,895	0,001	0,888	0,001
$\phi = 0.9$	0,737	0,001	0,711	0,000	0,720	0,001	0,737	0,000

Tabela 4. Potência e nível de significância empíricos do teste F, sobre a componente sazonal do modelo em que os erros são um processo AR(1) com ruído branco Normal(0, $\sigma = 20$) . T=40. Resultados para $\beta_0 = -100$, $\beta_0 = 0$, $\beta_0 = 100$, $\beta_0 = 500$.

As conclusões obtidas relativamente ao parâmetro β_0 são idênticas ao caso anterior, portanto, podemos afirmar que o teste F, no que respeita à significância da componente sazonal em séries com tendência, é robusto a variações dos parâmetros β_0 e β_1.

A leitura destas tabelas permite também constatar que um aumento do nível de correlação na série dos resíduos provoca uma diminuição na potência de teste e no nível de significância empírico. Na próxima secção analisa-se mais pormenorizadamente o comportamento do teste F para variações deste parâmetro.

6.1.2. Variações do Desvio Padrão da Série de Ruído Branco e da Dimensão da Série.

Ilustrado nos gráficos das séries temporais apresentadas neste capítulo (figuras 4 e 5), o desvio padrão da série a_t de ruído branco é um parâmetro que se prevê influenciar a *performance* do teste *F*, principalmente para amostras de menor dimensão. Em princípio, um incremento nesse parâmetro irá diminuir a eficácia da aplicação do teste *F*, em particular, na deteção da sazonalidade no modelo.

Para perceber o comportamento do teste F para variações de σ e para avaliar a influência do parâmetro de autocorrelação na sua *performance*, foram realizadas várias simulações de forma a cruzar alguns valores destes dois parâmetros. Os resultados obtidos para amostras de diferentes dimensões são apresentados na tabela 5.

		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.
		σ=10		σ=20		σ=30	
T=40	$\phi = 0$	1,000	0,058	0,989	0,058	0,810	0,042
	$\phi = 0.25$	1,000	0,027	0,991	0,028	0,780	0,031
	$\phi = 0.5$	1,000	0,007	0,982	0,010	0,662	0,007
	$\phi = 0.75$	1,000	0,003	0,903	0,002	0,435	0,001
	$\phi = 0.9$	0,999	0,000	0,724	0,000	0,276	0,000
		σ=10		σ=20		σ=30	
T=60	$\phi = 0$	1,000	0,050	1,000	0,042	0,944	0,059
	$\phi = 0.25$	1,000	0,022	1,000	0,017	0,932	0,022
	$\phi = 0.5$	1,000	0,008	0,999	0,007	0,884	0,006
	$\phi = 0.75$	1,000	0,000	0,987	0,001	0,656	0,001
	$\phi = 0.9$	1,000	0,000	0,862	0,000	0,352	0,000
		σ=10		σ=20		σ=30	
T=80	$\phi = 0$	1,000	0,053	1,000	0,041	0,991	0,050
	$\phi = 0.25$	1,000	0,021	1,000	0,018	0,987	0,027
	$\phi = 0.5$	1,000	0,004	1,000	0,005	0,972	0,004
	$\phi = 0.75$	1,000	0,000	1,000	0,000	0,849	0,000
	$\phi = 0.9$	1,000	0,000	0,943	0,000	0,465	0,000

Tabela 5. Potência e nível de significância empíricos do teste F, sobre a componente sazonal do modelo em que os erros são um processo AR(1) com ruído branco Normal. Resultados para T=40, T=60 e T=80; $\sigma = 10$, $\sigma = 20$ e $\sigma = 30$.

Analisando a tabela 5 é possível confirmar que a autocorrelação entre os erros afeta significativamente o nível de significância empírico, mesmo para valores não muito elevados do parâmetro autorregressivo. Embora o nível de significância decresça com o aumento da correlação dos erros, verifica-se que esta variação não é influenciada pela variância da série de ruído branco. Os resultados para T=40 estão ilustrados na figura 6.

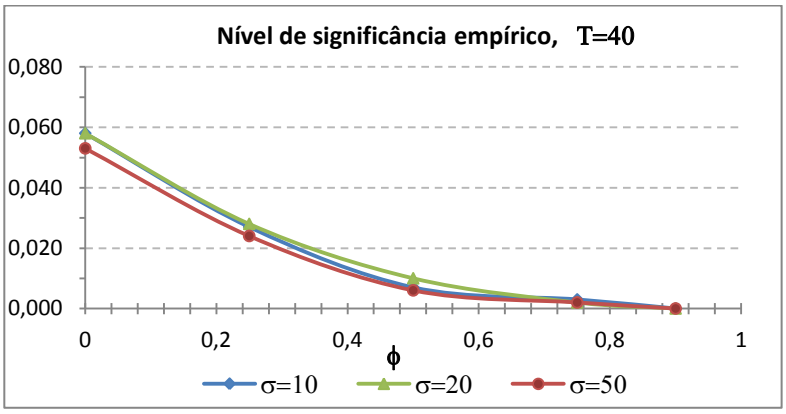


Figura 6. Nível de significância empírico do teste F para o modelo em que os erros são um processo autorregressivo AR(1) com ruído branco normal. T=40. Resultados para $\sigma = 10$, $\sigma = 20$ e $\sigma = 50$.

Como esperado, a potência de teste *F* é também afetada pela autocorrelação da série dos erros, verificando-se que, para variâncias mais elevadas, um incremento do parâmetro autorregressivo resulta numa redução da potência de teste. De referir que este comportamento também está relacionado com a diminuição do nível de significância empírico registado pois, quando existe correlação, a potência empírica de teste é calculada tendo por base um nível de significância que não corresponde ao fixado

inicialmente. Para se obter um valor de potência segundo um real nível de significância α, seria necessário ter conhecimento da verdadeira distribuição da estatística F, sob H_0, teoricamente desconhecida.

Analisando os resultados noutra perspetiva, constata-se que um aumento da variância da série de ruído branco conduz a uma quebra na potência empírica do teste, principalmente para os valores de autocorrelação mais elevados. Na figura 7 é apresentado um gráfico ilustrativo destes resultados para T=40.

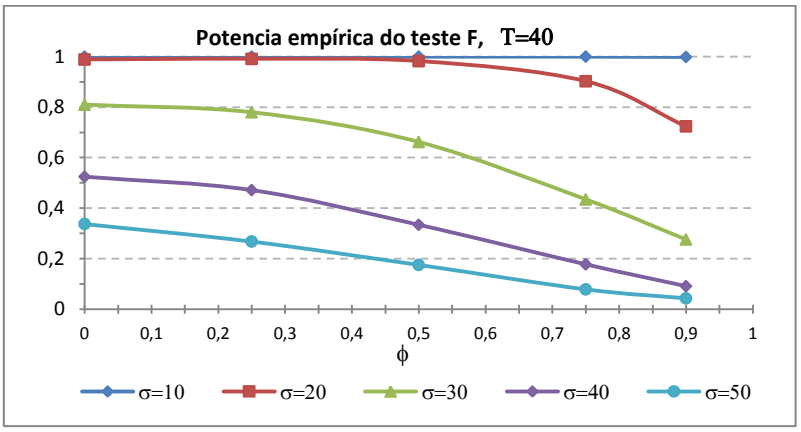


Figura 7. Potência empírica do teste F, para o modelo em que os erros são um processo autorregressivo AR(1) com ruído branco normal. T=40. Resultados para vários valores de σ.

De seguida, são comparados dois casos concretos, que exemplificam o comportamento do teste para um nível de correlação elevado face à diferença entre as variâncias de a_t. Por exemplo, para $\sigma = 10$ e $\phi = 0.9$ a aplicação do teste F às séries simuladas resulta num nível de significância empírico aproximadamente nulo, ou seja, o teste nunca ou quase nunca detetaria erradamente sazonalidade em séries simuladas sem sazonalidade. Além disso, a potência empírica registada foi de aproximadamente 1, isto é, o teste detetaria sempre ou quase sempre a sazonalidade em séries simuladas com componente sazonal. Por esta ordem de ideias, para o conjunto de valores utilizados, o teste F permitiria que se tomasse a decisão acertada, logo revelar-se-ia adequado.

Para $\sigma = 30$ e $\phi = 0.9$ o teste nunca ou quase nunca detetaria erradamente sazonalidade nas séries simuladas sem sazonalidade, contudo esta discrepância entre o nível de significância empírico e o nível de significância considerado inicialmente afeta consideravelmente a potência do teste, pois esta seria apenas aproximadamente 48%. Considerando o nível de significância a 5%, através de uma aproximação do quantil da distribuição F obtido pelo método de Monte Carlo, ter-se-ia obtido para a potência

empírica de teste um valor superior a 90% (este valor pode ser encontrado na tabela 6). Portanto, neste caso, o teste *F* mostrar-se-ia menos eficiente na deteção de sazonalidade no modelo.

Através da análise à tabela apresentada no anexo I, verifica-se que nas séries temporais com período P=12, o nível de significância empírico é menos afetado pela autocorrelação, face ao nominal, do que para P=4.

A dimensão da amostra é outro dos fatores que afeta o desempenho do teste F. Pode-se verificar que a potência empírica de teste é superior quando são simuladas séries de maiores dimensões. Repare-se que, mesmo para valores de correlação mais elevados, o aumento da dimensão da série melhorou substancialmente a potência de teste. Relativamente ao nível de significância empírico, este não é muito afetado por variações da dimensão da amostra. Na figura 8 é ilustrada a variação do nível de significância empírico face à dimensão da série simulada.

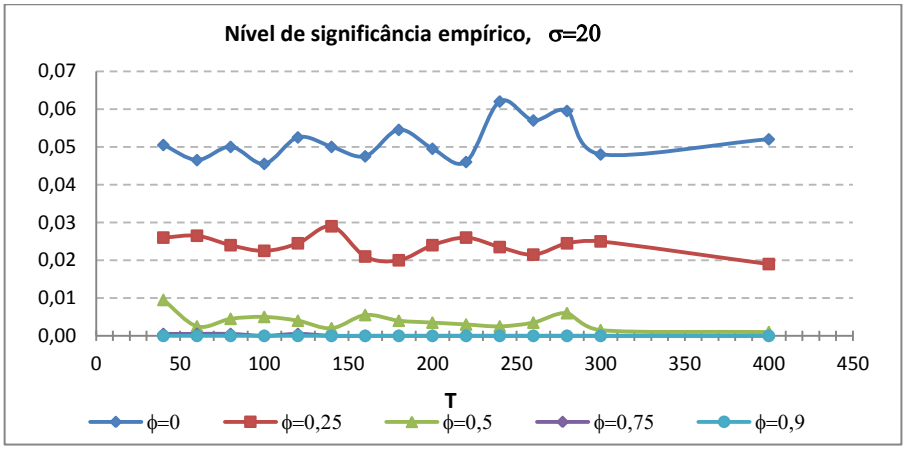


Figura 8. Nível de significância empírico do teste F para o modelo em que os erros são um processo autorregressivo AR(1) com ruído branco normal. Resultados para vários valores de T, $\sigma = 20$.

Procedendo a uma análise global aos resultados obtidos quando a série a_t é normalmente distribuída, reparamos que, apesar da autocorrelação nos erros afetar a *performance* do teste F usual, a probabilidade de deteção de sazonalidade numa série temporal com tendência não é, em geral, sobrestimada. Assim sendo, se o teste F usual indicar a rejeição da hipótese nula, esta decisão, em princípio, estará correta segundo o nível de significância nominal considerado. No entanto, o mesmo já não acontece na decisão de aceitação da hipótese nula.

6.1.3. Cálculo do Valor Crítico da Estatística F Usual através do Método de Monte Carlo

Nas análises anteriores constata-se que a presença de correlação na série dos erros afeta a distribuição da estatística de teste *F* usual sob H_0. No sentido de se ir ao encontro da real distribuição da estatística de teste quando existe autocorrelação, recorre-se neste estudo ao método de Monte Carlo para determinar uma melhor aproximação para o chamado valor crítico da distribuição de *F*, para um nível de significância α.

Na tabela 6 apresentam-se os resultados da potência e do nível de significância empíricos, obtidos segundo este método, tendo o valor $F_{crítico}$ sido calculado a partir de 2000 simulações e é tal que $P(F > F_{crítico}) \approx \alpha$.

		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.
		σ=10		σ=20		σ=30	
T=40	$\phi = 0$	1,000	0,053	0,994	0,053	0,803	0,053
	$\phi = 0.25$	1,000	0,051	0,995	0,033	0,846	0,054
	$\phi = 0.5$	1,000	0,053	0,998	0,055	0,883	0,038
	$\phi = 0.75$	1,000	0,049	1,000	0,046	0,918	0,048
	$\phi = 0.9$	1,000	0,047	0,998	0,050	0,913	0,056
		σ=10		σ=20		σ=30	
T=60	$\phi = 0$	1,000	0,047	1,000	0,043	0,951	0,060
	$\phi = 0.25$	1,000	0,048	1,000	0,061	0,970	0,041
	$\phi = 0.5$	1,000	0,054	1,000	0,042	0,975	0,044
	$\phi = 0.75$	1,000	0,056	1,000	0,070	0,994	0,051
	$\phi = 0.9$	1,000	0,061	1,000	0,054	0,986	0,053
		σ=10		σ=20		σ=30	
T=80	$\phi = 0$	1,000	0,049	1,000	0,049	0,988	0,055
	$\phi = 0.25$	1,000	0,059	1,000	0,043	0,994	0,054
	$\phi = 0.5$	1,000	0,071	1,000	0,049	0,998	0,049
	$\phi = 0.75$	1,000	0,053	1,000	0,052	1,000	0,050
	$\phi = 0.9$	1,000	0,050	1,000	0,052	1,000	0,063

Tabela 6. Potência e nível de significância empíricos obtidos a partir de uma aproximação do valor crítico da distribuição da estatística F através de simulações pelo método de Monte Carlo. Resultados para T=40, T=60 e T=80; $\sigma = 10$, $\sigma = 20$ e $\sigma = 30$.

Ao se estimar o valor crítico da distribuição da estatística de teste através do método de Monte Carlo, constata-se que a potência de teste aumenta com o incremento da autocorrelação e é superior à obtida através da estatística de teste *F* usual (quando $\phi > 0$). Contudo, do ponto de vista prático, este processo poderá não ser assim tão vantajoso pois, em situações normais, desconhecem-se, *a priori*, os parâmetros populacionais e a sua estimação conduzirá a erros que não foram contemplados nesta situação.

6.1.4. Incorporação da Matriz de Covariância na Estimação pelos Mínimos Quadrados

Segundo os resultados teóricos estudados, em dados que apresentam correlação serial significativa, quando é incorporada a matriz de covariância dos erros na estimação pelos MQG, a estatística F deverá produzir melhores resultados do que a estatística *F* usual, cujos coeficientes de regressão são estimados pelos MQO .

Para comparar o comportamento destas estatísticas na deteção de sazonalidade num modelo linear, para diferentes níveis de autocorrelação, foi realizado um estudo comparativo, através de simulação, em que foi calculado o nível de significância empírico e a potência empírica do teste. São também apresentados resultados da estatística F , em que na estimação pelos MQG é aplicada uma estimativa da matriz de covariância dos erros obtida a partir de $\hat{\phi}$. De referir que, neste caso, a estatística não tem uma distribuição F, pelo que é de esperar uma quebra de *performance*. Além disso, é teoricamente conhecido que a estimação de ϕ apresenta uma assimetria negativa acentuada, principalmente para amostras pequenas. Os resultados obtidos são apresentados na tabela 7.

		F_{usual}		F_{MQG} $(\hat{\phi} \neq \phi)$		$F_{MQG}{}^{*}$ $(\hat{\phi} = \phi)$	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.
σ=20	ϕ=0	0,993	0,044	0,990	0,059	0,993	0,044
	ϕ=0.25	0,991	0,029	0,995	0,061	0,998	0,054
	ϕ=0.5	0,979	0,009	0,998	0,037	1,000	0,053
	ϕ=0.75	0,905	0,001	1,000	0,040	1,000	0,050
	ϕ=0.9	0,753	0,000	1,000	0,028	1,000	0,048
		F_{usual}		F_{MQG} $(\hat{\phi} \neq \phi)$		$F_{MQG}{}^{*}$ $(\hat{\phi} = \phi)$	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.
σ=30	ϕ=0	0,792	0,046	0,824	0,053	0,787	0,051
	ϕ=0.25	0,763	0,028	0,820	0,052	0,829	0,052
	ϕ=0.5	0,679	0,009	0,866	0,044	0,881	0,052
	ϕ=0.75	0,441	0,002	0,927	0,031	0,947	0,054
	ϕ=0.9	0,266	0,000	0,946	0,027	0,975	0,049

Tabela 7. Potência e nível de significância empíricos do teste F para o modelo em que os erros são um processo AR(1) com ruído branco Normal. Comparação entre os valores obtidos segundo uma estimação pelos MQO e MQG. T=40 . Resultados para $\sigma = 20$ e $\sigma = 30$.

Como esperado, quando é utilizado o parâmetro populacional de autocorrelação no cálculo da estatística de teste, os resultados são bem melhores do que os obtidos

através da estatística de teste usual. Na estimação pelos MQG verificou-se que, para todos os níveis de correlação utilizados, o nível de significância empírico manteve-se próximo do valor estipulado inicialmente e que um incremento da correlação provoca um aumento na potência de teste, factos estes opostos aos registados aquando da aplicação da estatística F usual.

Quando o parâmetro autorregressivo é estimado, a *performance* do teste não é tão boa face à registada com a utilização do parâmetro populacional. Para valores mais elevados de autocorrelação, o nível de significância empírico sofre uma quebra e a potência de teste é ligeiramente inferior à de F_{MQG}. Todavia, continua-se a verificar que, para valores mais elevados de correlação, há um acréscimo na potência de teste.

Os resultados mostram ainda que, quando o parâmetro ϕ é estimado, a estatística F^*_{MQG} é, em geral, bem melhor que a estatística *F* usual, em que é desprezada a correlação dos dados. Para valores mais elevados de correlação, o nível de significância empírico, apesar de subestimado, é melhor e a potência de teste é bastante superior à obtida em *F* usual.

Relativamente às consequências de estimações incorretas da autocorrelação na estatística de teste, de referir o trabalho desenvolvido por Palm e Sneek (1984) , no qual concluíram que, sob H_0, a real distribuição de probabilidade de $\hat{F}$, numa regressão linear, tende a mover-se consideravelmente para a direita quando a autocorrelação é subestimada ($\hat{\phi} < \phi$), causando, portanto, um aumento na real probabilidade de se cometer um erro de tipo I. Por sua vez, quando ϕ é sobrestimado ($\hat{\phi} > \phi$), a distribuição move-se para a esquerda.

Para avaliar a influência que uma incorreta estimação do parâmetro autorregressivo poderá ter na *performance* da estatística de teste $\hat{F}$ na deteção de sazonalidade num modelo linear, foram realizadas várias simulações para vários valores de ϕ e $\hat{\phi}$. Os resultados obtidos podem ser visualizados na tabela 8. As simulações foram realizadas para T=40.

		$\hat{\phi}=0$		$\hat{\phi}=0.25$		$\hat{\phi}=0.5$		$\hat{\phi}=0.75$		$\hat{\phi}=0.9$	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível Signif.	Potência	Nível signif.
σ=20	ϕ=0	0,990	0,053	0,992	0,076	0,992	0,114	0,992	0,127	0,991	0,112
	ϕ=0.25	0,992	0,026	0,995	0,044	0,999	0,089	0,997	0,105	0,999	0,107
	ϕ=0.5	0,982	0,007	0,996	0,025	0,998	0,047	0,999	0,078	1,000	0,105
	ϕ=0.75	0,894	0,003	0,983	0,003	0,999	0,014	1,000	0,050	1,000	0,058
	ϕ=0.9	0,718	0,001	0,918	0,003	0,996	0,010	1,000	0,027	1,000	0,041
		$\hat{\phi}=0$		$\hat{\phi}=0.25$		$\hat{\phi}=0.5$		$\hat{\phi}=0.75$		$\hat{\phi}=0.9$	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível Signif.	Potência	Nível signif.
σ=30	ϕ=0	0,792	0,054	0,807	0,071	0,824	0,097	0,812	0,110	0,818	0,119
	ϕ=0.25	0,779	0,030	0,837	0,048	0,854	0,080	0,866	0,117	0,889	0,114
	ϕ=0.5	0,679	0,010	0,794	0,024	0,883	0,050	0,937	0,076	0,925	0,088
	ϕ=0.75	0,441	0,001	0,674	0,007	0,878	0,020	0,955	0,055	0,962	0,068
	ϕ=0.9	0,264	0,000	0,531	0,003	0,805	0,007	0,952	0,034	0,979	0,049

Tabela 8. Potência e nível de significância empíricos do teste F, com estimação pelos MQG, para o modelo em que os erros são um processo AR(1), quando a estimativa $\hat{\phi}$ afasta-se do verdadeiro parâmetro populacional ϕ. T=40 . Resultados para $\sigma=20$ e $\sigma=30$.

Os resultados obtidos mostram que, quando aplicado à componente sazonal de um modelo linear, o teste *F* apresenta um comportamento diferente do observado nos modelos lineares testados por Palm e Sneek. De facto, no teste à sazonalidade, quando $\hat{\phi}>\phi$, a real probabilidade de se cometer um erro de tipo I tende a aumentar. Por sua vez, se ocorrer uma subestimação acentuada de ϕ, o nível de significância empírico reduzirá substancialmente. Segundo Palm e Sneek, os resultados de aplicação do teste podem variar de acordo com os regressores utilizados, pelo que esta poderá ser apontada como uma das causas das discrepâncias verificadas entre os dois estudos.

Verificamos ainda que a potência de teste aumenta ligeiramente com a sobrestimação do parâmetro autorregressivo, contudo caso este seja subestimado, a potência de teste poderá vir a ser bastante afetada, decaindo substancialmente nos casos em que o erro de estimação de ϕ é maior.

Pela análise dos resultados obtidos e dado que ϕ tende a ser subestimado, podemos concluir que uma estimação inapropriada do parâmetro autorregressivo irá na maioria das vezes causar uma perca de eficácia do teste *F* na deteção de sazonalidade. É preciso ter em conta que este resultado não é regra e que será necessário alguma prudência na análise de resultados quando houver suspeitas que a autocorrelação possa ter sido incorretamente estimada.

6.2. Processo AR(1) em que os a_t tem distribuição Gama

Com a finalidade de incorporar alguma assimetria na componente dos erros aleatórios, considerou-se a_t com distribuição Gama com parâmetro de forma r e parâmetro de escala α. Assim, de forma similar ao estudo realizado para o caso da normalidade da série a_t, procurou-se estudar a *performance* do teste F aplicado à componente sazonal de uma série temporal, para diferentes níveis de autocorrelação dos erros aleatórios, quando estes são gerados por um processo AR(1).

As séries a_t foram geradas para um conjunto fixo de parâmetros de forma r, enquanto o parâmetro de escala α variou em função de um conjunto de valores de variância da distribuição estipulados inicialmente. Os coeficientes sazonais usados na simulação foram os mesmos que os utilizados em 6.1.

Na tabela 9 são apresentados alguns dos resultados obtidos nas simulações para os vários níveis de autocorrelação da série dos erros considerados. De salientar que foi efetuada uma translação da série a_t de forma a garantir que o seu valor médio fosse nulo. Considerou-se a dimensão da amostra T=40.

		r=0.5		*r*=1		*r*=1.5		*r*=3		*r*=4		N(0, σ=10)	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível Signif.
σ=10	ϕ=0	1,000	0,039	1,000	0,034	1,000	0,048	1,000	0,050	1,000	0,045	1,000	0,058
	ϕ=0.25	1,000	0,024	1,000	0,025	1,000	0,026	1,000	0,025	1,000	0,024	1,000	0,027
	ϕ=0.5	1,000	0,004	1,000	0,007	1,000	0,006	1,000	0,009	1,000	0,006	1,000	0,007
	ϕ=0.75	1,000	0,001	0,999	0,000	1,000	0,001	1,000	0,000	1,000	0,002	1,000	0,003
	ϕ=0.9	0,997	0,001	0,997	0,001	0,997	0,000	0,999	0,000	0,999	0,000	0,999	0,000
		r=0.5		*r*=1		*r*=1.5		*r*=3		*r*=4		N(0, σ=20)	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível Signif.
σ=20	ϕ=0	0,964	0,035	0,975	0,036	0,983	0,044	0,984	0,044	0,982	0,053	0,989	0,058
	ϕ=0.25	0,970	0,014	0,974	0,019	0,983	0,030	0,983	0,025	0,988	0,029	0,991	0,028
	ϕ=0.5	0,950	0,005	0,961	0,006	0,960	0,006	0,968	0,006	0,972	0,006	0,982	0,010
	ϕ=0.75	0,872	0,001	0,874	0,002	0,882	0,001	0,894	0,001	0,893	0,001	0,903	0,002
	ϕ=0.9	0,753	0,000	0,730	0,001	0,757	0,000	0,736	0,000	0,743	0,000	0,724	0,000
		r=0.5		*r*=1		*r*=1.5		*r*=3		*r*=4		N(0, σ=30)	
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.	Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível Signif.
σ=30	ϕ=0	0,795	0,035	0,805	0,043	0,797	0,041	0,798	0,043	0,796	0,051	0,810	0,042
	ϕ=0.25	0,782	0,015	0,786	0,027	0,769	0,021	0,747	0,026	0,764	0,020	0,780	0,031
	ϕ=0.5	0,723	0,004	0,703	0,007	0,674	0,006	0,679	0,008	0,671	0,007	0,662	0,007
	ϕ=0.75	0,510	0,001	0,475	0,001	0,471	0,000	0,442	0,001	0,453	0,001	0,435	0,001
	ϕ=0.9	0,376	0,000	0,308	0,000	0,302	0,000	0,282	0,002	0,272	0,001	0,276	0,000

Tabela 9. Potência e nível de significância empíricos do teste F, para o modelo em que os erros são um processo AR(1) com ruído branco Gama(*r*, α) . T=40. Resultados para $\sigma = 10$, $\sigma = 20$ e $\sigma = 30$; *r*=0.5, *r*=1, *r*=1.5, *r*=3 e *r*=4.

É possível verificar que, para os parâmetros de forma mais usuais ($r \geq 1{,}5$), a assimetria introduzida na série de ruído branco não provocou alterações muito significativas na *performance* do teste *F*, face ao observado aquando da normalidade de a_t. Para parâmetros de forma mais reduzidos, em que a distribuição apresenta uma assimetria semelhante ou mais elevada que a da exponencial (r=1), o nível de significância empírico é mais afetado. Pela análise da tabela e por outras simulações realizadas que não constam dos resultados apresentados, concluiu-se que um aumento da assimetria da série de ruído branco, isto é, uma diminuição do parâmetro de forma, tem como consequência uma redução no nível de significância empírico. Relativamente à potência do teste, não se detetaram alterações dignas de registo em relação ao obtido com a distribuição normal.

7. Estudo de Séries Reais

7.1. Número de Dormidas nos Açores em Hotelaria Tradicional e Turismo de Espaço Rural

O Arquipélago dos Açores, situado no meio do Atlântico Norte, a cerca de mil milhas a oeste de Portugal continental, foi eleito pela revista *National Geographic* como o 8º melhor destino para o verão de 2011. A contribuir para esta distinção está a sua paisagem natural, recortada por montanhas vulcânicas de onde sobressaem as crateras e lagoas que explicam a origem destas ilhas, convencendo muitas vezes os visitantes a optarem por alojamento em espaços tradicionais ou de turismo rural.

O seu clima temperado marítimo, com temperaturas médias amenas ao longo de todo o ano e com precipitação abundante, sobretudo no outono e inverno, é um dos fatores que fomenta um padrão de sazonalidade, quer na procura global deste destino turístico, quer nos alojamentos. Neste trabalho iremos cingir o nosso estudo ao comportamento no tempo da variável dormidas nos Açores em hotelaria tradicional e turismo rural.

Em primeiro lugar, pretendemos ajustar a variável em estudo a um modelo matemático e averiguar se existe um padrão de sazonalidade que seja significativo de ser incluído no modelo. Para o efeito utilizou-se os valores referentes ao número de dormidas em espaços de turismo rural e hotelaria tradicional na Região Autónoma dos Açores, desde Janeiro de 2003 até Dezembro de 2010 (fonte: Serviço Regional de Estatística dos Açores). A série temporal em estudo é ilustrada na figura 9.

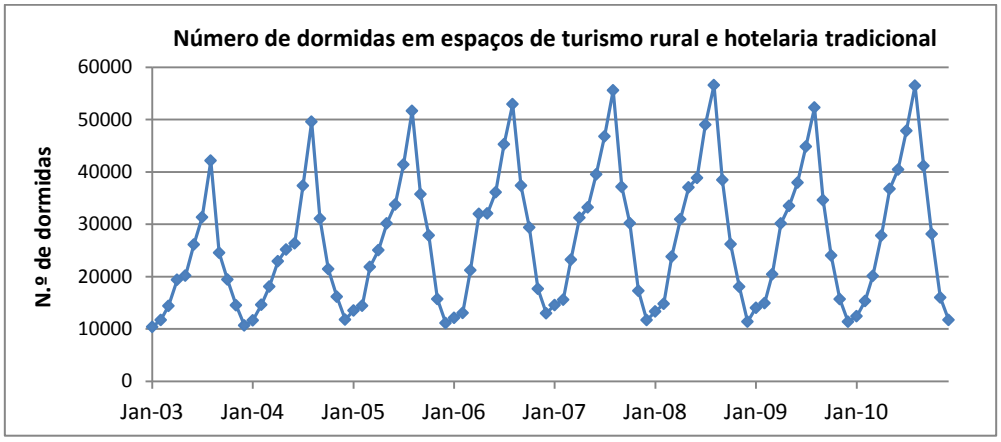


Figura 9. Série da variável número de dormidas em espaços de turismo rural e hotelaria tradicional nos Açores. Valores observados entre janeiro 2003 e dezembro 2010.

O gráfico desta série mostra uma tendência que, aparentemente, é ligeiramente crescente, acompanhada de um efeito sazonal. Depois de ajustado o conjunto de

observações a um modelo de regressão linear com tendência e sazonalidade, aplicou-se o teste F para avaliar a significância da componente sazonal. Por fim é feita uma apreciação sobre a fiabilidade dos resultados obtidos na aplicação do teste F a esta série temporal, tendo em conta quer os resultados do teste quer as conclusões obtidas ao longo deste trabalho.

- Estimação de tendência e sazonalidade:

Uma vez que a unidade temporal é o mês e que o período total a que correspondem os dados é de 96 semanas, o modelo em estudo admite a representação linear:

$$Z_t = \beta_0 + \beta_1 t + \sum_{i=1}^{12} x_{ti} S_i + \varepsilon_t ,$$

em que *t* representa o mês e x_{ti} são neste caso dadas por:

$$x_{ti} = \begin{cases} 1 \quad se \quad t = 12(j-1)+i, \text{ para algum } j = 1,...,8 \\ 0, \quad \text{caso contrário} \end{cases} .$$

Atendendo à condição imposta aos coeficientes sazonais, é necessário reparametrizar o modelo anterior para que seja possível a estimação pelo método dos mínimos quadrados, como por exemplo:

$$Z_t = \beta_1 t + \sum_{i=1}^{12} x_{ti} S_i + \varepsilon_t ,$$

com $\beta_0 = \overline{S}$ e $s_i = S_i - \overline{S}$.

Os valores calculados para estes parâmetros são apresentados na tabela 10.

Coeficientes		valores
Termo independente	β_0	22.546,19
Declive	β_1	93,19
Janeiro	s_1	-13.806,61
Fevereiro	s_2	-12.345,42
Março	s_3	-6.347,11
Abril	s_4	609,71
Maio	s_5	4.089,02
Junho	s_6	7.874,83
Julho	s_7	15.878,90
Agosto	s_8	24.966,21
Setembro	s_9	7.722,90
Outubro	s_{10}	-1.551,29
Novembro	s_{11}	-11.101,23
Dezembro	s_{12}	-15.990

Tabela 10. Estimativas dos coeficientes do modelo linear da série número de dormidas em espaços de turismo rural e hotelaria tradicional, pelo método dos mínimos quadrados.

A partir dos resultados obtidos, representou-se graficamente o modelo ajustado à série original, tal como pode ser observado na figura 10.

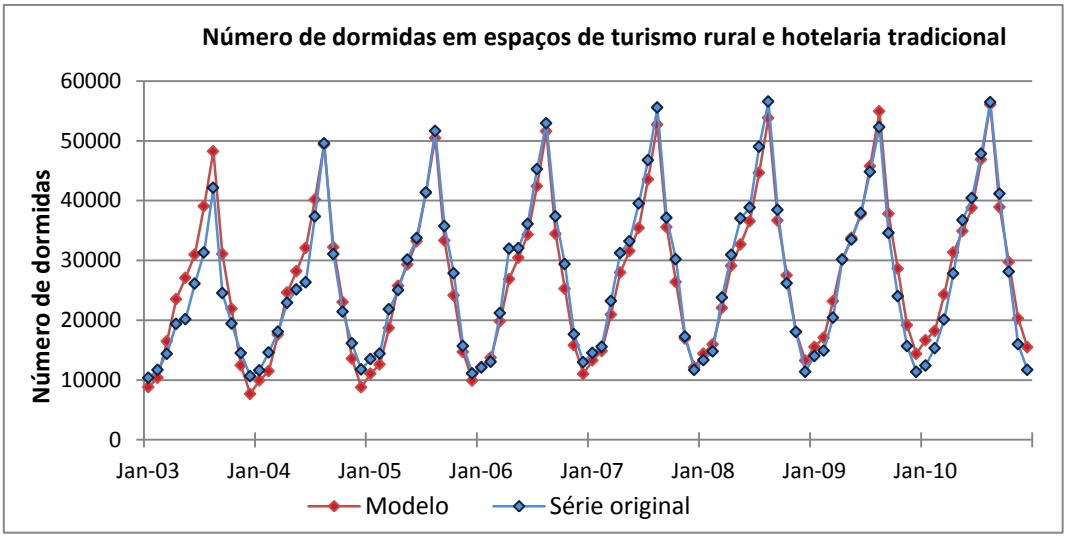


Figura 10. Modelo ajustado da série da variável número de dormidas em espaços de turismo rural e hotelaria tradicional resultante da estimação pelos mínimos quadrados.

Como podemos visualizar, este modelo ajusta-se bem à série em estudo, tal como pode ser comprovado pelo valor de R^2= 0.9431, bastante próximo de 1.

Para avaliar a significância da componente sazonal desta regressão linear, aplicou-se a estatística *F* usual para testar o seguinte teste de hipóteses:

$$H_0 : s_1 = s_2 = ... = s_{12} = 0 \qquad Vs \qquad H_1 : \exists i, i = 1,...,12 : s_i \neq 0$$

O cálculo da estatística de teste resultou no valor *F*=132.22, que para os graus de liberdade do numerador e denominador, 11 e 83 respetivamente, corresponde a um *valor-p* aproximadamente nulo. Conclui-se, então, que a componente sazonal é significativa para o modelo.

Apesar destes resultados parecerem bastante convincentes, a validade destas conclusões está dependente de alguns pressupostos, nomeadamente a normalidade dos resíduos, a variância constante e a independência dos erros.

Primeiramente será averiguada a existência de correlação na série dos resíduos, ou seja, na série sem tendência e sem sazonalidade. O gráfico desta série pode ser observado na figura 11.

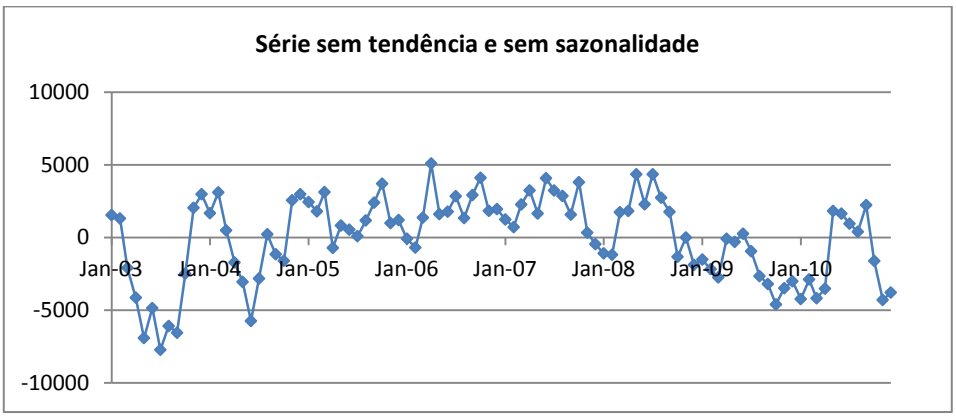


Figura 11. Série residual (vs tempo) resultante do ajustamento pelo método dos mínimos quadrados.

Para a deteção de autocorrelação na série dos resíduos e, em caso de existência, para a identificação do processo subjacente, recorreu-se à função de autocorrelação (FAC) e à função de autocorrelação parcial (FACP), estimadas a partir da série sem tendência e sem sazonalidade. Calcularam-se e examinaram-se estas funções amostrais para os primeiros 24 valores. Os valores calculados encontram-se na tabela 10 e estão representados na figura 12.

Intervalo	FAC	FACP		Intervalo	FAC	FACP
1	0,754	0,754		**13**	0,359	0,081
2	0,551	-0,042		**14**	0,199	-0,229
3	0,359	-0,099		**15**	0,073	-0,039
4	0,188	-0,088		**16**	-0,049	-0,055
5	0,033	-0,104		**17**	-0,112	0,005
6	0,009	0,165		**18**	-0,115	0,065
7	0,022	0,063		**19**	-0,076	-0,052
8	0,106	0,158		**20**	-0,039	-0,051
9	0,243	0,200		**21**	0,001	-0,099
10	0,314	-0,027		**22**	0,024	-0,094
11	0,364	0,078		**23**	0,012	-0,010
12	0,372	0,021		**24**	-0,017	-0,038

Tabela 10. Valores da FAC e FACP da série residual, obtidos após ajustamento pelo método dos mínimos quadrados.

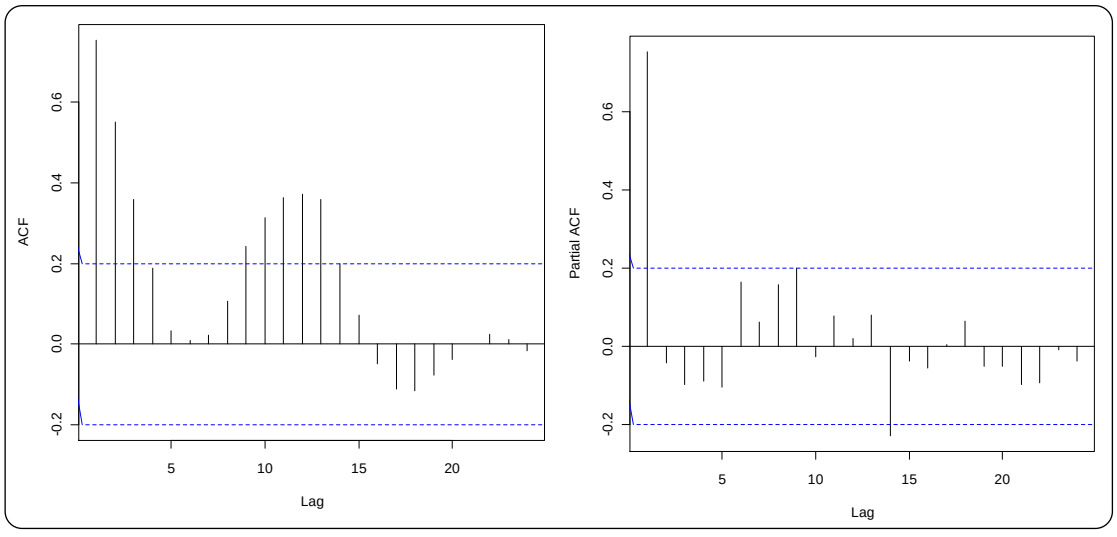


Figura 12. Gráficos da FAC e FACP da série residual, obtidos após ajustamento pelo método dos mínimos quadrados.

Por um lado, Podemos observar que a FAC converge gradualmente para zero e que a FACP apresenta uma quebra brusca entre o primeiro e o segundo valor. Por outro lado, na função de autocorrelação parcial, para um espaçamento superior a um, todos os valores pertencem ao intervalo $\pm 2/\sqrt{96}$, à exceção do intervalo 14, cujo valor correspondente se encontra muito próximo do extremo esquerdo do intervalo de confiança referido. Assim sendo, podemos concluir que a série dos resíduos se ajusta a um processo autorregressivo de 1ª ordem, AR(1), caracterizado pelo seguinte modelo:

$$e_t = \phi e_{t-1} + a_t ,$$

onde $\{a_t\}$ é um processo de ruído branco e ϕ uma constante real.

Para a identificação do processo estar completa resta apenas estimar o parâmetro ϕ associado ao modelo. Para tal recorreu-se às equações de Yule-Walquer, que, no caso do modelo AR(1), traduzem-se em:

$$\hat{\phi} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \hat{\rho}_1$$

Consultando a tabela 10 obtém-se $\hat{\phi} = 0{,}754$.

Em seguida, foi avaliada a qualidade do modelo, ou seja, foi analisado se os resíduos resultantes do ajustamento ao modelo AR(1) se comportam como um processo de ruído branco.

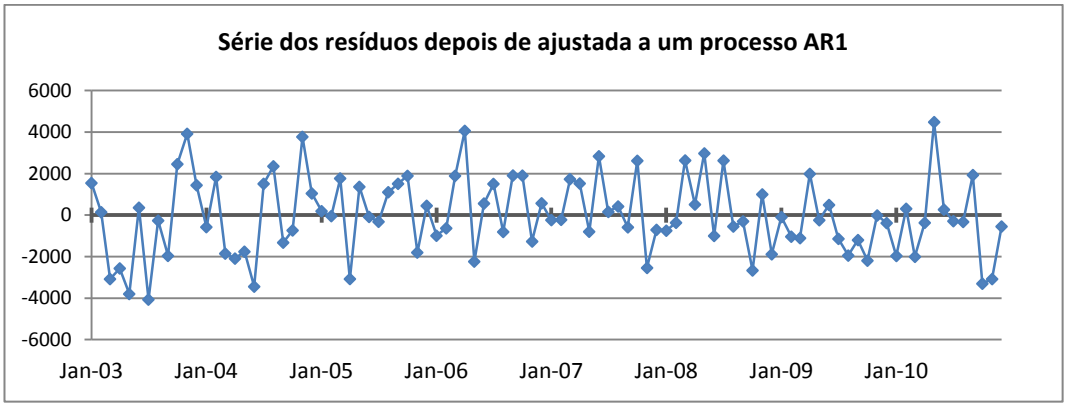


Figura 13. Série dos resíduos depois de ajustada a um processo AR(1) através da transformação $a_t = e_t - \hat{\phi} e_{t-1}$.

Por observação da figura 13 fica-se com a ideia que, de facto, a variância e a média são constantes. Contudo, apesar dos resíduos parecerem homogéneos, para se confirmar a hipótese de que os resíduos se comportam como um ruído branco, realizou-se o seguinte teste de hipóteses:

$$H_0 : \rho_1 = ... = \rho_{24} = 0 \quad \text{Vs} \quad H_1 : \rho_j \neq 0 \ , \quad j = 1,...,24$$

Aplicou-se um dos teste de Portmanteau, mais precisamente o teste de Ljung-Box (1978), cuja estatística de teste é dada por:

$$Q = T(T+2)\sum_{k=1}^{s} \rho_k^2 / (T-k) .$$

Obteve-se o valor Q=32.8824. Como $\chi^2_{0,95}(23) = 35.17246$, para um nível de significância de 5% não rejeitamos a hipótese nula e conclui-se que os resíduos são não correlacionados.

Podemos ainda testar a normalidade dos resíduos depois de ajustada a série a um processo AR(1). Para tal recorreu-se ao teste de Saphiro-Wilk, cujo valor da estatística de teste, W=0.9884, corresponde a um *valor-p* de 0.5651. Assim, para os níveis de significância usuais, conclui-se que estes resíduos são normalmente distribuídos.

Através dos resultados anteriores, podemos então admitir que, depois de ajustados a um processo AR(1), os resíduos seguem um processo de ruído branco. Assim sendo, apesar de não se verificarem todos os pressupostos de aplicabilidade do teste *F* usual, devemos aceitar a decisão que deste advém. Nos estudos de simulação, concluiu-se que sendo os resíduos gerados por um processo AR(1), se o teste F indicar a rejeição de H_0, então essa decisão deverá ser aceite pois o nível de significância empírico é, em geral, subestimado na presença de autocorrelação.

Foi ainda aplicada a estatística de teste F, na qual é incorporada a matriz de covariância na estimação pelos mínimos quadrados. Como estimativa do parâmetro autorregressivo, utilizou-se o valor $\hat{\phi} = 0{,}754$, todavia este valor poderá não ser uma boa estimativa pois, para além do erro natural proveniente de estimação, o estimador utilizado não é centrado. Aplicada a estatística de teste, obteve-se o valor F =128.44, que, como seria de esperar, resulta numa decisão idêntica à do teste F usual.

7.2. Exportações de Automóveis em Portugal

Foi realizado um segundo estudo em séries reais, neste caso, respeitante a uma variável na área da indústria. O estudo diz respeito às exportações de automóveis para transporte de passageiros, de Portugal para o resto do mundo, por trimestre, compreendidas entre Janeiro de 2000 e Dezembro de 2010 (fonte: INE).

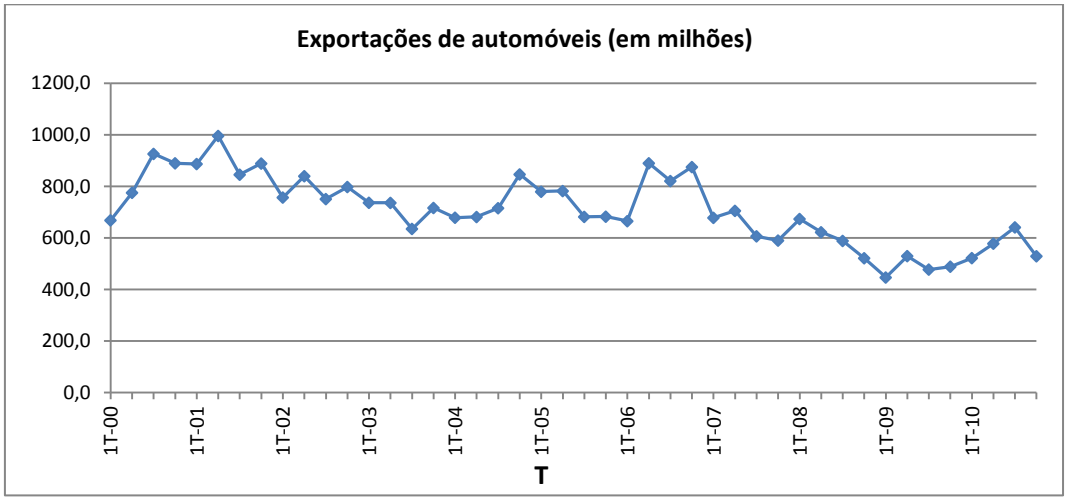


Figura 14. Série da variável exportações de automóveis em Portugal. Valores observados entre janeiro 2000 e dezembro 2010.

De forma análoga ao exemplo anterior, os dados foram ajustados a um modelo linear com tendência e sazonalidade e posteriormente aplicou-se o teste *F* para avaliar a qualidade desse ajustamento.

Procedendo a uma análise visual dos dados ilustrados na figura 14, notamos uma tendência linear decrescente, no entanto, não é identificada a ideia de sazonalidade. Procedeu-se então a um ajustamento a um modelo através de uma regressão linar.

Neste caso, a unidade temporal é o trimestre, num período total de 11 anos, o que resulta no seguinte modelo linear, já reparametrizado, para permitir a estimação pelos mínimos quadrados:

$$Z_t = \beta_1 t + \sum_{i=1}^{4} x_{ti} S_i + \varepsilon_t ,$$

em que *t* representa o mês e x_{ti} são, neste caso, dadas por

$$x_{ti} = \begin{cases} 1 & se \quad t = 4(j-1)+i, \text{ para algum } j = 1,...,11 \\ 0, & \text{caso contrário} \end{cases},$$

e com $\beta_0 = \overline{S}$ e $s_i = S_i - \overline{S}$.

Os valores calculados para estes parâmetros são apresentados na tabela 11.

Coeficientes		valores
Termo independente	β_0	887,05
Declive	β_1	-7,97
1 Trimestre	s_1	-38,64
2 Trimestre	s_2	27,78
3 Trimestre	s_3	-4,78
4 Trimestre	s_4	15,63

Tabela 11. Estimativas dos coeficientes do modelo linear da série exportações de automóveis, pelo método dos mínimos quadrados.

A partir dos resultados obtidos, representou-se graficamente o modelo ajustado à série original, tal como pode ser observado na figura 15.

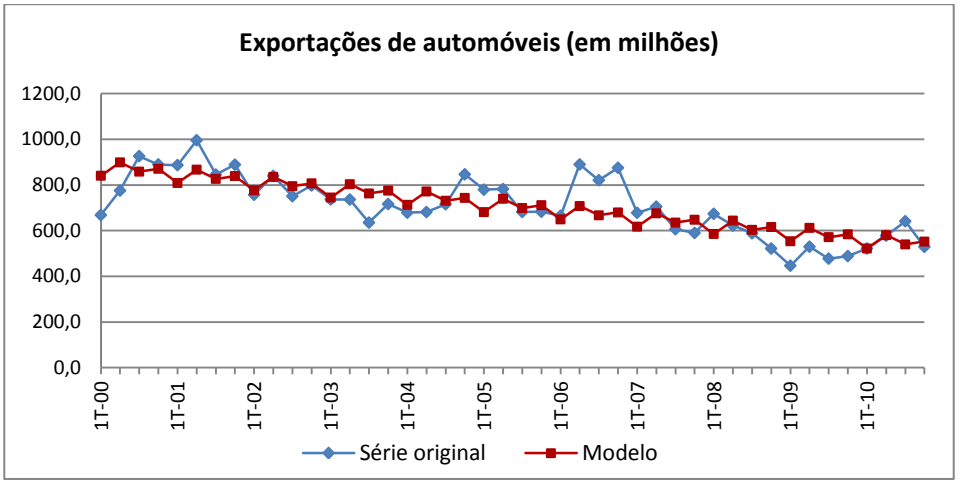


Figura 15. Modelo ajustado da série da variável exportações de automóveis, resultante da estimação pelos mínimos quadrados.

A componente de tendência parece ajustar-se bem à série original. Procedendo ao teste F usual aplicado ao modelo completo, obtemos o valor de F=15.01 para a estatística de teste, que corresponde a um *valor-p* aproximadamente nulo. Podemos então afirmar que o ajuste da série ao modelo completo é significativo. Relativamente à significância da componente sazonal do modelo, obteve-se para a estatística de teste um valor de F= 1.1908, que para os graus de liberdade 3 (numerador) e 39 (denominador), corresponde a um *valor-p* 0.3258. Concluímos assim que a componente sazonal não é significativa para o modelo.

É preciso ter novamente em conta que a validade destas conclusões está dependente da verificação dos pressupostos de aplicabilidade do teste F. Neste sentido, foi efetuado um estudo série dos resíduos para verificação de independência, normalidade e homogeneidade de variâncias. O gráfico desta série pode ser observado na figura 16.

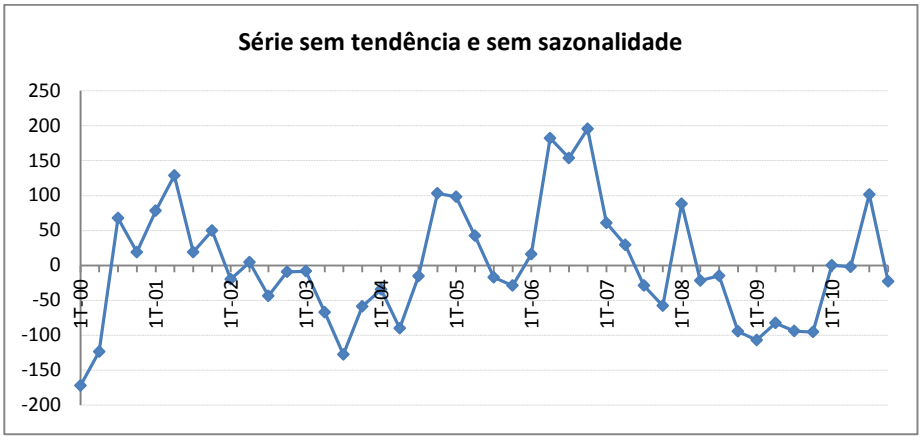


Figura 16. Série residual (vs tempo) resultante do ajustamento pelo método dos mínimos quadrados.

A partir desta série calcularam-se as estimativas das funções de autocorrelação (FAC) e de autocorrelação parcial (FACP) para os primeiros 16 valores. Os resultados obtidos são apresentados na tabela 12 e os respetivos gráficos na figura 17.

Intervalo	FAC	FACP
1	1,000	0,550
2	0,550	-0,053
3	0,265	-0,136
4	0,029	-0,085
5	-0,110	0,103
6	-0,061	0,008
7	-0,017	-0,015
8	0,014	-0,178
9	-0,092	-0,233
10	-0,274	-0,101
11	-0,351	-0,233
12	-0,444	-0,057
13	-0,345	0,095
14	-0,118	-0,023
15	0,009	-0,099
16	0,039	0,043

Tabela 12. Estimativas dos coeficientes do modelo linear da série exportações de automóveis, pelo método dos mínimos quadrados.

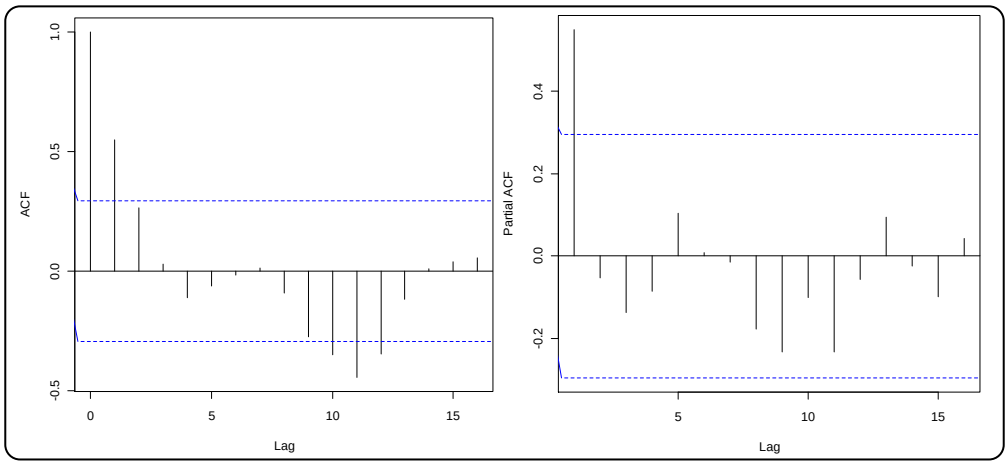


Figura 17. Gráficos da FAC e FACP da série residual, obtidos após ajustamento pelo método dos mínimos quadrados.

Analisando estas duas funções, constata-se novamente que a FAC converge gradualmente para zero e que a FACP apresenta uma quebra brusca entre o primeiro e o segundo valor e, para um espaçamento superior a um, todos os valores pertencem ao intervalo $\pm 2/\sqrt{44}$. Conclui-se então que a série dos resíduos se ajusta a um processo autorregressivo de 1ª ordem, AR(1).

Consultando a tabela 12, obtém-se uma estimativa para o parâmetro autorregressivo, $\hat{\phi} = 0{,}550$.

De seguida, verificou-se se os resíduos a_t, resultantes do ajustamento ao modelo AR(1) se comportam, de facto, como um processo de ruído branco. Para averiguar se a correlação entre esses valores era nula, aplicou-se o teste de Ljung-Box (1978) no seguinte teste de hipóteses:

$$H_0 : \rho_1 = ... = \rho_{16} = 0 \quad \text{Vs} \quad H_1 : \rho_j \neq 0 \ , \quad j = 1,...,16$$

Obteve-se o valor Q=12.8751, o que, para 15 graus de liberdade, corresponde a um *valor-p* de 0.6119. Não existem razões para rejeitar a hipótese nula, logo conclui-se que os resíduos são não correlacionados.

Por aplicação do teste de Saphiro-Wilk, aos níveis de significância usuais, conclui-se que a série a_t é normalmente distribuída. Obteve-se uma estatística de teste de W=0,9645, que corresponde a um *valor-p* de 0.5651.

Através dos resultados anteriores, pode-se então admitir que, depois de ajustados a um processo AR(1), os resíduos a_t seguem um processo de ruído branco. Assim sendo, a componente estocástica (dos erros) encontra-se sob as condições que foram consideradas no estudo por simulação.

Foi ainda aplicada a estatística de teste *F* pelos MQG, na qual é incorporada a estimativa da matriz de covariância. Como estimativa do parâmetro autorregressivo, utilizou-se o valor $\hat{\phi} = 0.550$. Obteve-se o valor *F*= 3.3585, que para os graus de liberdade 3 (numerador) e 39 (denominador), corresponde a um *valor-p* 0.028. A um nível de significância de 5% dever-se-ia rejeitar a hipótese inicial e concluir que a componente sazonal é significativa no modelo.

A aplicação das estatísticas de teste F pelos MQO e pelos MQG conduziu a decisões opostas no que diz respeito à significância da componente sazonal do modelo. Atendendo às conclusões obtidas no estudo por simulação, verificou-se que a estatística *F* pelos MQG, na qual é considerada uma estimativa da matriz de covariância, é mais eficaz na deteção de sazonalidade que a estatística F usual. Contudo, é preciso ter em conta que uma sobrestimação de ϕ, embora menos provável, resulta num aumento da probabilidade de se rejeitar incorretamente a hipótese nula. Assim, tendo em conta estes resultados e os valores provenientes das duas estatísticas de teste, a um nível de significância de 5%, os resultados obtidos no que respeita à significância da componente sazonal na série temporal relativa às exportações de automóveis, não conduzem a uma decisão fiável, segundo o nível de significância nominal considerado.

8. Discussão de Resultados e Investigação Futura

Uma das principais conclusões deste trabalho é que a correlação nos resíduos influencia seriamente a distribuição da estatística de teste F usual no teste à significância da componente sazonal em séries temporais com tendência linear, principalmente para valores de correlação mais elevados e é influenciada pela estimação do parâmetro autorregressivo. Este resultado vai, em parte, de encontro a outros resultados que foram estudados, para situações análogas, para outras variações dos parâmetros

Os estudos de simulação mostraram que um aumento da autocorrelação nos resíduos, para além da perda de potência do teste, provoca uma subestimação do nível de significância empírico. A partir destes factos, podemos inferir, que se a um determinado nível de significância, a aplicação do teste *F* usual indicar uma presença significativa de sazonalidade no modelo, esta decisão deverá ser aceite pois, na presença de autocorrelação, tende a verificar-se uma deslocação para a direita do valor calculado da estatística de teste, continuando este a ser superior ao valor crítico da distribuição *F*. Por outro lado, os resultados da aplicação do teste poderão não ser fiáveis se estes indicarem que a componente sazonal não é significativa, sendo, nestes casos, preferível recorrer-se a outros métodos estatísticos alternativos.

Relativamente à utilização da estatística de teste *F* em que a autocorrelação nos resíduos é tida em conta através da estimação pelos mínimos quadrados generalizados, constata-se que os resultados obtidos são, em geral, melhores do que na utilização dos mínimos quadrados ordinários. Contudo é preciso ter em conta que a necessidade de estimação do parâmetro autorregressivo levará a que a *performance* do teste seja afetada, face ao caso em que é conhecido o verdadeiro valor populacional. Devido ao enviesamento à esquerda do estimador utilizado, o parâmetro autorregressivo tende a ser subestimado, o que provoca uma diminuição no nível de significância empírico e na potência empírica do teste. Resultados contrários registam-se quando o parâmetro autorregressivo é sobrestimado.

Outra das conclusões obtidas é que a estatística de teste *F* para testar a significância da componente sazonal, não parece ser afetada por variações dos restantes coeficientes do modelo linear, nomeadamente de tendência e da ordenada na origem, mesmo na presença de autocorrelação nos resíduos. O nível de significância empírico também parece não ser muito influenciado pela dimensão da amostra considerada nem

pela variância da série de ruído branco. Todas estas conclusões devem ser apresentadas com alguma precaução, e tendo em conta os intervalos de valores que foram considerados nos estudos de simulação, não podendo ser generalizadas precipitadamente.

Na presença de uma considerável autocorrelação dos resíduos é preferível incorporar-se esta componente na estatística de teste *F* através de processos como a estimação pelos MQG ou efetuar-se uma transformação de variáveis antes da aplicação do teste *F* usual. Quando a autocorrelação aparentar ser pouco significativa, pode continuar-se a usar a estatística de teste F usual, tendo sempre o cuidado de se analisar melhor quando da aplicação do teste resultar que a componente sazonal seja pouco significativa, pois a probabilidade de se rejeitar a sazonalidade é subestimada na presença de autocorrelação.

O estudo realizado mostrou ainda que os resultados de aplicação do teste F parecem não ser muito afetados por desvios à normalidade da série de ruído branco, desde que estes não sejam demasiado acentuados. Experimentado com recurso à distribuição gama, verificou-se que para uma assimetria desta distribuição igual ou superior à da distribuição exponencial, os valores do nível de significância empírico e da potência empírica situaram-se relativamente próximos dos obtidos aquando da distribuição normal. Quando a assimetria é mais acentuada, estes dois valores sofrem um decréscimo face ao obtido na presença de normalidade.

Numa análise futura, seria pertinente estender o estudo da performance do teste *F* na significância da componente sazonal em séries cujos erros são outros processos estacionários e averiguar se o comportamento do teste é semelhante ao registado pelo processo AR(1) estudado neste trabalho.

Bibliografia e Referências

Alpuim, T., 2003. *Séries Temporais*. Associação dos Estudantes da Faculdade de Ciências da Universidade de Lisboa.

Bratley, P., **Fox**, B. L., **Scharge**, L. E., 1987. *A Guide to Simulation*, 2nd Ed., Springer Verlag.

Cochrane, D., **Orcutt**, G.H., 1949. *Application of Least Squares regression to relationships containing autocorrelated error terms*. Journal of American Statistical Association. 44, 32-41.

Faraway, J., 2002. *Pratical Regression and Anova using R.*

Jobson, J., **Fienberg**, E., **Olkin**,I., 1991. *Applied Multivariate Data Analisys Volume 1: Regression and Experimental Design*, Springer – Verlag.

Maroco, J., 2007. *Análise Estatística com a utilização do SPSS,* Ed. Sílabo, Lisboa.

Palm, F., **Sneek**, J., 1984. *Significance tests and spurious correlation in regression models with autocorrelated errors*. Statistische Hefte. 25, 87-105.

Ramos. M. R., 2006**.** *Testes de tendência com Aplicação à Avaliação da Qualidade da Água.* Tese de Doutoramento, Universidade de Lisboa.

Rao, C. R., **Toutenburg**, H., 1999. *Linear Models Leasts Squares and Alternatives*, 2nd Ed., Springer.

Robert, C., **Casella**, G., 2010. *Introducing Monte Carlo Methods with R*, Springer.

Seber, G., **Lee**. A., 2003. *Linear Regression Analysis*, 2nd Ed., John Wiley & Sons Publication.

Sen, A., **Srivastava**, M., 1990. *Regression analysis: theory, methods and applications.* Springer-Verlag.

Shumway, R., **Stoffer**, D., 2011. *Time Series Analysis and Its Applications*, 3rd Ed., Springer New York Dordrecht Heidelberg London.

Verzani, J., 2004. *Using R for Introductory Statistics*, Chapman & Hall / CRC.

Yan, X. **Su**, X., 2009. *Linear Regression Analysis : Theory and Computing*, World Scientific Publishing.

http://cran.r-project.org/doc/contrib/Torgo-ProgrammingIntro.pdf

http://cran.r-project.org/doc/manuals/

-

Anexo 1

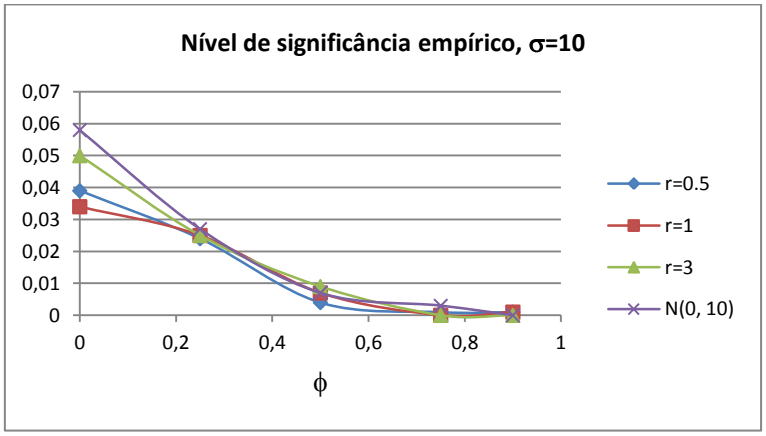
Tabelas e gráficos com os resultados complementares para o nível de significância e potência empírica de teste F, referentes ao capítulo 6 – Estudos de simulação.

6.1. Processo AR(1) em que os a_t têm distribuição normal

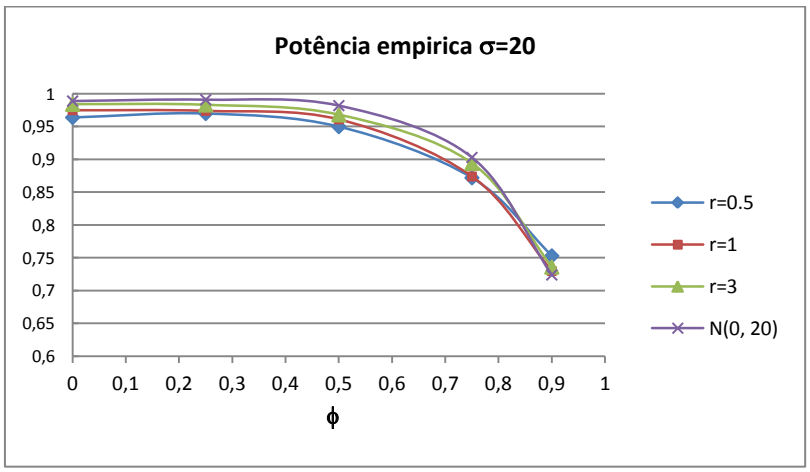
		Potência	Nível Signif.	Potência	Nível signif.	Potência	Nível signif.
		σ=10		σ=20		σ=30	
T=120	$\phi = 0$	1,000	0,047	0,997	0,059	0,848	0,044
	$\phi = 0.25$	1,000	0,039	0,988	0,045	0,736	0,036
	$\phi = 0.5$	1,000	0,045	0,937	0,037	0,580	0,040
	$\phi = 0.75$	1,000	0,015	0,650	0,012	0,262	0,016
	$\phi = 0.9$	0,961	0,002	0,251	0,001	0,058	0,002

Nível de significância e potência empírica de teste *F*. P=12. Potência calculada para simulações em que s_i=(-13.8,-12.3,-6.3,0.6,4.1,7.9,15.9,25,7.7,-1.6,-11.1,-16) .

6.2. Processo AR(1) em que os a_t têm distribuição Gama



Nível de significância empírico do teste *F* para P=4. Resultados para σ=10.



Potência empírica do teste *F* para P=4 e σ=20.

Anexo 2

Código utilizado no R para a componente de simulação

Teste F pelos MQO e pelos MQG em que ε_i é um processo AR(1) com distribuição Normal

```
##########Função para simular uma sequência de n elementos segundo um processo
AR1(phi=ro)
simula_AR1<-function(n,ro,media,desviop)
{
        #simulação dos resíduos anteriores a Z0 - 20 residuos
        at<-rnorm(20,media,desviop)  #simulação da série at
        Z<-vector() #criação da variável dos resíduos
        Z[1]<-at[1]
        for (t in 2:20)
                {Z[t]<-ro*Z[t-1]+at[t]}
        Z0<-Z[20]

        #simulação da série dos resíduos de dimensão n
        at<-rnorm(n,media,desviop)  #simulação da série at
        Z<-vector() #criação da variável dos resíduos
        Z[1]<-Z0
        for (t in 2:n)
                {Z[t]<-ro*Z[t-1]+at[t]}  #Calculo Zt segundo um processo AR1 (a partir de
Zt-1 e at)
Z #devolução da série dos resíduos Zt
}
####### FIM FUNÇAO simular AR1


#####Função para criar uma SERIE TEMPORAL com tendência linear e sazonalidade.
Serietemporal<-function(beta1,Si,Zt)
{
        P<-length(Si) #período da componente sazonal
        xti<-vector() #criação das variáveis indicatrizes xti
        for (contador in 1:(T/P)) xti<-rbind(diag(P),xti) #criação da matriz xti a partir
da repetição da matriz identidade(P)
        t<-1:T #definição da variável independente t de dimensão T
        y<-beta1*t+xti%*%Si+Zt # Definição da série temporal com tendência linear e
sazonalidade.
y #devolução da série temporal simulada
}

####### FIM FUNÇÃO SERIE TEMPORAL


### Função para Executar o TESTE F a uma serie temporal com at NORMAL
TesteF_normal<-function(T,Si,beta1,phi,media,dp)
{
        Zt<-simula_AR1(T,phi,media,dp) #criação dum processo AR1
        y<-Serietemporal(beta1,Si,Zt)   #criação de uma série temporal com tendência e
sazonalidade
        t<-1:T
        P<-length(Si)
        xti<-vector()
        for (contador in 1:(T/P)) xti<-rbind(diag(P),xti)   #criação da matriz xti a
partir da repetição da matriz identidade(P)
        X<-cbind(t,xti)#Criação da matriz X das variávis independentes

        b<-solve((t(X)%*%X)) %*% (t(X)%*%y)   ###Aplicação dos MQO para determinar
estimativas dos coeficientes
        modelo_estimado<-X%*%b
        residuos<-y-modelo_estimado

        ### TESTE à significância do modelo completo
        TSS<-sum((y-mean(y))^2)  #variabilidade total
```

```
        RSSc<-sum((y-modelo_estimado)^2)  #parte da variabilidade explicada pelos
resíduos
        regSS<-TSS-RSSc       #parte da variabilidade explicada pela regressão
        F1<-(regSS/P)/(RSSc/(T-P-1)) #estatistica de teste F usual
        valorp<-1-pf(F1,P,T-P-1)
        ### FIM TESTE à significância do modelo completo

        ### Teste à significância da componente sazonal(modelo reduzido)
        X2<-cbind(rep(1,T),t)   #matriz das variáveis independentes do modelo reduzido
        bi<-solve((t(X2)%*%X2)) %*% (t(X2)%*%y) #aplicação dos MQO ao modelo reduzido
        modelo_estimado<-X2%*%bi
        RSSi<-sum((y-modelo_estimado)^2)
        F2<-((RSSi-RSSc)/(P-1))/(RSSc/(T-P-1)) #estatística de teste
        valorp_r<-1-pf(F2,P-1,T-P-1)

        ####  Teste F pelos mínimos quadrados generalizados - COM AUTOCORRELAÇÃO:
        #criação da matriz das restrições C depois de eliminado Sp
        C<-diag(P)
        C<-cbind(0, C) #inserção da coluna correspondente a beta1
        C<-cbind(0, C) #inserção da coluna correspondente a beta0
        C<-C[,-(P+2)]  #eliminação da coluna relativa Sp
        C<-C[-P,]      #eliminação da última linha (coeficientes nulos)

        #matriz X sem Sp
        xti<-vector()
        xti<-rbind(diag(P))
        xti<-xti[-P,]
        xti<-rbind(xti,-1)
        X<-xti
        for (contador in 2:(T/P)) X<-rbind(X,xti)   #criação da matriz xti a partir da
repetição de xti
        X<-cbind(matrix(c(1:T)),X) #Inserção da coluna referente à variável t
        X<-cbind(1,X)  #Inserção da coluna referente a beta0
        X<-X[,-(P+2)]   #matriz X sem Betap

        #Matriz de covariancia - OMEGA
        b<-solve((t(X)%*%X)) %*% (t(X)%*%y)   #calculo das estimativas b pelos mínimos
quadrados ordinários
        residuos<-y-X%*%b
        phi_e<-pacf(residuos)[[1]][1]  #calculo da estimativa de phi.
        omega<-diag(T)
        #peenchimento da triangular superior de omega
        for (linha in 1:T)
                {contador<-0
                for (coluna in linha:T)
                        {omega[linha,coluna]<-phi_e^contador
                        contador<-contador+1}}

        #peenchimento da triangular inferior
        for (coluna in 1:T)
                {contador<-0
                for (linha in coluna:T)
                        {omega[linha,coluna]<-phi_e^contador
                        contador<-contador+1}}

        r<-rep(0,P-1) #matriz r em R*Beta=r
        R<-C
        k<-dim(X)[2]   #número de restrições a testar (característica de X)
        m<-dim(C)[1]   #número de coeficientes a testar

        ##Cálculo da estatística de teste
        K<-R%*%solve(t(X)%*%solve(omega)%*%X)%*%t(X)%*%solve(omega)
        Q<-solve(R%*%solve(t(X)%*%solve(omega)%*%X)%*%t(R))/m
        M<-(solve(omega)-
solve(omega)%*%X%*%solve(t(X)%*%solve(omega)%*%X)%*%t(X)%*%solve(omega))/(T-k)
        Fcorr<-(t(K%*%y-r)%*%Q%*%(K%*%y-r))  /   (t(y)%*%M%*%y)

        #FIM Teste F com AUTOCORRELAÇÃO
```

```
        #Variável Resumo da aplicaçao do TESte
        F<-list(completo=F1, Fcorr=Fcorr, reduzido=F2, pvalue=valorp_r, serie=y,
beta1=b[1], beta0=mean(b[-1]), coefSaz=b[-1], res=residuos, phi_e=phi_e )
        F
}

####### Função para determinar a POTENCIA EMPÍRICA DO TESTE F (com at NORMAL)
PotenciaF_normal<-function(T,Si,beta1,phi,nsim,media,dp,alfa)
{
        contagem<-0
        for (contador in 1:nsim) if
(TesteF_normal(T,Si,beta1,phi,media,dp)$reduzido>qf(0.95,P-1,T-P-1)) contagem<-
contagem+1
        Potencia<-contagem/nsim
        print(Potencia)
}

### Função para determinar a POTENCIA DO TESTE F - COM CORRELAÇAO
PotenciaF_normal_corr<-function(T,Si,beta1,phi,nsim,media,dp,alfa)
{
        contagem<-0
        for (contador in 1:nsim) if
(TesteF_normal(T,Si,beta1,phi,media,dp)$Fcorr>qf(0.95,P-1,T-P-1)) contagem<-contagem+1
        Potencia<-contagem/nsim
        print(Potencia)
}

##### Função para determinar a POTENCIA DO TESTE F - COM E SEM CORRELAÇAO: series iguais
PotenciaF_normal_corr<-function(T,Si,beta1,phi,nsim,media,dp,alfa)
{
        contagem<-0
        contagemc<-0
        for (contador in 1:nsim)
        {
                TesteF<-TesteF_normal(T,Si,beta1,phi,media,dp)
                if (TesteF$reduzido>qf(0.95,P-1,T-P-1)) contagem<-contagem+1
                if (TesteF$Fcorr>qf(0.95,P-1,T-P-1)) contagemc<-contagemc+1
        }
        Potencia<-contagem/nsim
        Potenciac<-contagemc/nsim
        Resultado<-cbind(Potenciac,Potencia)
        print(Resultado)
}

### Função para determinar a POTENCIA DO Teste F ao MODELO COMPLETO -
PotenciaF_normal_completo<-function(T,Si,beta1,phi,nsim,media,dp,alfa)
{
contagem<-0
        for (contador in 1:nsim) if
(TesteF_normal(T,Si,beta1,phi,media,dp)$Fcompleto>qf(0.95,P,T-P-1)) contagem<-contagem+1
        Potencia<-contagem/nsim
        print(Potencia)
}
#Exemplo do cálculo da potência empírica do teste F para uma série de período 4.
# definição dos parâmetros iniciais
si<-c(-26,3,20,3)
beta0<-177
beta1<-0.7
phi<-0
dp<-10
T<-40
P<-4
nsim<-2000
alfa<-0.05
Si<-si+beta0
media<-0
PotenciaF_normal(T,Si,beta1,phi,nsim,media,dp,alfa) #execução
```

Teste F usual em que ε_i é um processo AR(1) com distribuição Gama

```
##########Função para simular uma sequência de n elementos segundo um processo
AR1(phi=ro)
simula_AR1<-function(n,ro,shape,scale)
{
	#simulação dos resíduos anteriores a Z0 - 20 residuos
	at<-rgamma(20,shape,scale=scale) #simulação da série at
	at<-at-shape*scale
	Z<-vector() #criação da variável dos resíduos
	Z[1]<-at[1]
	for (t in 2:20)
		{Z[t]<-ro*Z[t-1]+at[t]}
	Z0<-Z[20]

	#simulação da série dos resíduos de dimensão n
	at<-rgamma(n,shape,scale=scale)  #simulação da série at
	at<-at-shape*scale #translaçao da serie de ruido branco para se obter uma serie
de ruidos com media 0
	Z<-vector() #criação da variável dos resíduos
	Z[1]<-Z0
	for (t in 2:n)
		{Z[t]<-ro*Z[t-1]+at[t]}  #Calculo Zt segundo um processo AR1 (a partir de
Zt-1 e at)
Z #devolução da série dos resíduos Zt
}
####### FIM FUNÇAO simular AR1


#####Função para criar uma SERIE TEMPORAL com tendência linear e sazonalidade.
Serietemporal<-function(beta1,Si,Zt)
{
	P<-length(Si) #período da componente sazonal
	xti<-vector() #criação das variáveis indicatrizes xti
	for (contador in 1:(T/P)) xti<-rbind(diag(P),xti) #criação da matriz xti a partir
da repetição da matriz identidade(P)
	t<-1:T #definição da variável independente t de dimensão T
	y<-beta1*t+xti%*%Si+Zt # Definição da série temporal com tendência linear e
sazonalidade.
y #devolução da série temporal simulada
}

####### FIM FUNÇÃO SERIE TEMPORAL



### Função para Executar o TESTE F a uma serie temporal com at GAMA
TesteF_gamma<-function(T,Si,beta1,phi,shape,scale)
{
	Zt<-simula_AR1(T,phi,shape,scale) #criação dum processo AR1
	y<-Serietemporal(beta1,Si,Zt)   #criação de uma série temporal com tendência e
sazonalidade
	t<-1:T
	P<-length(Si)
	xti<-vector()
	for (contador in 1:(T/P)) xti<-rbind(diag(P),xti)   #criação da matriz xti a
partir da repetição da matriz identidade(P)
	X<-cbind(t,xti)#Criação da matriz X das variávis independentes

	b<-solve((t(X)%*%X)) %*% (t(X)%*%y)   ###Aplicação dos MQO para determinar
estimativas dos coeficientes
	modelo_estimado<-X%*%b
	residuos<-y-modelo_estimado

	### TESTE à significância do modelo completo
	TSS<-sum((y-mean(y))^2)  #variabilidade total
```

```
        RSSc<-sum((y-modelo_estimado)^2)  #parte da variabilidade explicada pelos
resíduos
        regSS<-TSS-RSSc       #parte da variabilidade explicada pela regressão
        F1<-(regSS/P)/(RSSc/(T-P-1)) #estatistica de teste F usual
        valorp<-1-pf(F1,P,T-P-1)
        ### FIM TESTE à significância do modelo completo

        ### Teste à significância da componente sazonal(modelo reduzido)
        X2<-cbind(rep(1,T),t)   #matriz das variáveis independentes do modelo reduzido
        bi<-solve((t(X2)%*%X2)) %*% (t(X2)%*%y) #aplicação dos MQO ao modelo reduzido
        modelo_estimado<-X2%*%bi
        RSSi<-sum((y-modelo_estimado)^2)
        F2<-((RSSi-RSSc)/(P-1))/(RSSc/(T-P-1)) #estatística de teste
        valorp_r<-1-pf(F2,P-1,T-P-1)


        #Variável Resumo da aplicaçao do TESte
        F<-list(completo=F1, reduzido=F2, pvalue=valorp_r, serie=y, beta1=b[1],
beta0=mean(b[-1]), coefSaz=b[-1], res=residuos)
        F

}

####### Função para determinar a POTENCIA EMPÍRICA DO TESTE F (com at GAMA)
PotenciaF_gamma<-function(T,Si,beta1,phi,nsim,shape,scale,alfa)
{
        contagem<-0
        for (contador in 1:nsim) if
(TesteF_gamma(T,Si,beta1,phi,shape,scale)$reduzido>qf(0.95,P-1,T-P-1)) contagem<-
contagem+1
        Potencia<-contagem/nsim
        print(Potencia)
}

### Função para determinar a POTENCIA DO Teste F ao MODELO COMPLETO -
PotenciaF_gamma_completo<-function(T,Si,beta1,phi,nsim,shape,scale,alfa)
{
        contagem<-0
        for (contador in 1:nsim) if
(TesteF_gamma(T,Si,beta1,phi,shape,scale)$Fcompleto>qf(0.95,P,T-P-1)) contagem<-
contagem+1
        Potencia<-contagem/nsim
        print(Potencia)
}

#Exemplo do cálculo da potência empírica do teste F para uma série de período 4.
# definição dos parâmetros iniciais
si<-c(-26,3,20,3)
beta0<-177
beta1<-0.7
phi<-0
shape<-1
scale<-10
T<-40
P<-4
nsim<-2000
alfa<-0.05
Si<-si+beta0
media<-0

PotenciaF_gamma (T,Si,beta1,phi,nsim,shape,scale,alfa)  #execução
```

```
##########Função para simular uma sequência de n elementos segundo um processo
AR1(phi=ro)
simula_AR1<-function(n,ro,media,desviop)
{
       #simulação dos resíduos anteriores a Z0 - 20 residuos
       at<-rnorm(20,media,desviop)  #simulação da série at
       Z<-vector() #criação da variável dos resíduos
       Z[1]<-at[1]
       for (t in 2:20)
              {Z[t]<-ro*Z[t-1]+at[t]}
       Z0<-Z[20]

       #simulação da série dos resíduos de dimensão n
       at<-rnorm(n,media,desviop)  #simulação da série at
       Z<-vector() #criação da variável dos resíduos
       Z[1]<-Z0
       for (t in 2:n)
              {Z[t]<-ro*Z[t-1]+at[t]}  #Calculo Zt segundo um processo AR1 (a partir de
Zt-1 e at)
Z #devolução da série dos resíduos Zt
}
####### FIM FUNÇAO simular AR1


#####Função para criar uma SERIE TEMPORAL com tendência linear e sazonalidade.
Serietemporal<-function(beta1,Si,Zt)
{
       P<-length(Si) #período da componente sazonal
       xti<-vector() #criação das variáveis indicatrizes xti
       for (contador in 1:(T/P)) xti<-rbind(diag(P),xti) #criação da matriz xti a partir
da repetição da matriz identidade(P)
       t<-1:T #definição da variável independente t de dimensão T
       y<-beta1*t+xti%*%Si+Zt # Definição da série temporal com tendência linear e
sazonalidade.
y #devolução da série temporal simulada
}

####### FIM FUNÇÃO SERIE TEMPORAL


### Função para Executar o TESTE F a uma serie temporal com at NORMAL - COM CORRELAÇÃO
TesteF_normal_corr<-function(T,Si,beta1,phi,media,dp,phi_e)
{
      Zt<-simula_AR1(T,phi,media,dp)
      y<-Serietemporal(beta1,Si,Zt)
      t<-1:T
      P<-length(Si)

      #matriz C das restrições lineares
      C<-diag(P)
      C[P,]<--1
      C<-cbind(0, C)
      C<-cbind(0, C)
      C<-C[,-(P+2)]
      C<-C[-P,]  #é necessário eliminar 2 linhas correspondentes a b0 e t

      #matriz X sem sp
      xti<-vector()
      xti<-rbind(diag(P))
      xti<-xti[-P,]
      xti<-rbind(xti,-1)
      X<-xti
      for (contador in 2:(T/P)) X<-rbind(X,xti)   #criar a matriz xti a partir da
repetição de xti
      X<-cbind(matrix(c(1:T)),X)
      X<-cbind(1,X)  #matriz com coluna de beta0
      X<-X[,-(P+2)]   #matriz sem beta12
```

```
      #Matriz de covariancia - OMEGA

      omega<-diag(T)
      #peenchimento da triangular superior
      for (linha in 1:T)
               {contador<-0
               for (coluna in linha:T)
                      {omega[linha,coluna]<-phi_e^contador
                      contador<-contador+1}}

      #peenchimento da triangular inferior
      for (coluna in 1:T)
               {contador<-0
                      for (linha in coluna:T)
                      {omega[linha,coluna]<-phi_e^contador
                      contador<-contador+1}}

      r<-rep(0,P-1) # 2ª membro das equações das restrições lineares
      R<-C
      k<-dim(X)[2]   #dimensão da matrix X - colunas
      m<-dim(C)[1]

      b<-solve((t(X)%*%solve(omega)%*%X)) %*% (t(X)%*%solve(omega)%*%y)   #estimativas
dos coeficientes de regressão
      residuos<-X%*%b

      #Calculo da estatística de teste
      K<-R%*%solve(t(X)%*%solve(omega)%*%X)%*%t(X)%*%solve(omega)
      Q<-solve(R%*%solve(t(X)%*%solve(omega)%*%X)%*%t(R))/m
      M<-(solve(omega)-
solve(omega)%*%X%*%solve(t(X)%*%solve(omega)%*%X)%*%t(X)%*%solve(omega))/(T-k)
      Fcorr<-(t(K%*%y-r)%*%Q%*%(K%*%y-r))  /   (t(y)%*%M%*%y)


      #Resumo da aplicaçao do TESte
      F<-list(Fcorr=Fcorr, serie=y, beta1=b[2], beta0=b[1], coefSaz=b[-1][-1],
res=residuos, phi_e=phi_e )
      F

}


### FUNÇAO PARA DETERMINAR A POTENCIA DO TESTE F - COM CORRELAÇAO
PotenciaF_normal_corr<-function(T,Si,beta1,phi,nsim,media,dp,alfa,phi_e)
      {
      contagem<-0
      for (contador in 1:nsim)
         {
         TesteF<-TesteF_normal_corr(T,Si,beta1,phi,media,dp,phi_e)
         if (TesteF$Fcorr>qf(0.95,P-1,T-P-1)) contagem<-contagem+1
         }
      Potencia<-contagem/nsim*100
      print(Potencia)
}


#Exemplo do cálculo da potência empírica do teste F para uma série de período 4.
# definição dos parâmetros iniciais
si<-c(-26,3,20,3)
beta0<-177
beta1<-0.7
phi<-0
dp<-10
T<-40
P<-4
nsim<-2000
alfa<-0.05
Si<-si+beta0
media<-0
phi_e<-0.9

PotenciaF_normal_corr(T,Si,beta1,phi,nsim,media,dp,alfa,phi_e) #execução
```

```
### FUNÇAO PARA DETERMINAR A POTENCIA ”da estatística F”, EM QUE at é NORMAL
PotenciaF_normal_crit<-function (T,Si,beta1,phi,nsim,media,dp,alfa)
{
      # Cálculo do quantil(0.95) da estatística F usual sob H0 (FCritico).
      estatistica_teste<-vector()
      contagem<-0
      for (contador in 1:nsim)
           {
           TesteF<-TesteF_normal(T,rep(0,P)+beta0,beta1,phi,media,dp)
           estatistica_teste[contador]<-TesteF$reduzido
      }
      estatistica_teste<-sort(estatistica_teste)  #ordenar a distribuição
      Fcritico<-estatistica_teste[0.95*nsim]    #encontrar o valor F crítico


##Cálculo da potência da “estatística F” -  necessita da função definida acima
contagem<-0
for (contador in 1:nsim)
      {
      TesteF<-TesteF_normal(T,Si,beta1,phi,media,dp)
      if (TesteF$reduzido>Fcritico) contagem<-contagem+1
      }
      Potencia<-contagem/nsim
      print(Potencia)
}
```

Errata: Tese de Mestrado

Teste *F* na Regressão Linear Múltipla para Dados Temporais com Correlação Serial

Pág	Linha	Onde se lê	Deve ler-se
2	1	10 capítulos	8 capítulos
2	5	5, 6, 7	5, 6, 7 e 8
9	14	$\beta_0, \beta_1, \beta_2, ..., \beta_p$	$\beta_0, \beta_1, \beta_2, ..., \beta_k$
9	17	$\beta_j, j = 1, ..., p$	$\beta_j, j = 1, ..., k$
13	7	A partir daqui deduz-se o estimador centrado σ^2: $E(s^2) = \frac{\mathbf{e^T e}}{n-k-1} = \frac{\sigma^2(n-k-1)}{n-k-1} = \sigma^2.$	A partir daqui deduz-se o estimador centrado de σ^2: $s^2 = \frac{\mathbf{e^T e}}{n-k-1},$ com $E(s^2) = \frac{\sigma^2(n-k-1)}{n-k-1} = \sigma^2.$
10	7	$Y = X\beta + \varepsilon,$ em que $\varepsilon_i \quad \cap N(0, \sigma^2 \cdot I_n)$	$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon},$ em que $\boldsymbol{\varepsilon} \cap \mathbf{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_n)$
13	14	$H_0 : \beta_0 = \beta_1 = \beta_2 = ... = \beta_k = 0$ vs $H_1 : \exists j : \beta_j \neq 0, \quad j = 0, ..., k$	$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ vs $H_1 : \exists j : \beta_j \neq 0, \quad j = 1, ..., k$
13	15	Utilizando a notação matricial, as hipóteses a testar assumem o seguinte aspeto: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$	(eliminar)
15	1	$F = \frac{\mathbf{b^T X^T X b}/k}{\left[\mathbf{Y^T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}\right]/n-k-1}$.	$F = \frac{\mathbf{Y^T}\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T} - \mathbf{J}/n\right)\mathbf{Y}^2/k}{\mathbf{Y^T}\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right]\mathbf{Y}/(n-k-1)},$ com $\mathbf{J} = \mathbf{1} \cdot \mathbf{1^T}$ em que $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$
15	9	$\delta = \frac{\beta^T X^T X \beta}{\sigma^2}$	$\delta = \frac{\boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}(\mathbf{I} - \mathbf{J}/n)\mathbf{X}\boldsymbol{\beta}}{\sigma^2}$
18	4	$F = \frac{(\mathbf{Rb} - \mathbf{r})^{\mathbf{T}}\left[\mathbf{C}\left(\mathbf{X^T X}\right)^{-1}\mathbf{C^T}\right]^{-1}(\mathbf{Rb} - \mathbf{r})}{ms^2} \sim F_{m,n-k},$	$F = \frac{(\mathbf{Rb} - \mathbf{r})^{\mathbf{T}}\left[\mathbf{C}\left(\mathbf{X^T X}\right)^{-1}\mathbf{C^T}\right]^{-1}(\mathbf{Rb} - \mathbf{r})}{ms^2} \sim F_{m,n-k-1},$
19	20	$F = \frac{\left[\mathbf{Y^T}\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T I} - \mathbf{X}_2\left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\right)\mathbf{T}\right]/(p-1)}{\left[\mathbf{Y^T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}\right]/(n-p-1)} \sim F_{p-1,n-p-1}$	$F = \frac{\left[\mathbf{Y^T}\left(\mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T I} - \mathbf{X}_2\left(\mathbf{X}_2^{\mathbf{T}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathbf{T}}\right)\mathbf{Y}\right]/(p-1)}{\left[\mathbf{Y^T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}\right)\mathbf{Y}\right]/(n-p-1)} \sim F_{p-1,n-p-1}$
53	24	*performance* da estatística de teste $\hat{F}$	*performance* da estatística de teste *F* pelos MQG
58	8	96 semanas	96 meses
63	15	t representa o mês	t representa o trimestre