INSTITUTO UNIVERSITÁRIO DE LISBOA

2020/2021

Licenciatura em Ciência de Dados – Pós-Laboral

2º Ano, 1º Semestre

Introdução a Modelos Dinâmicos

**Grupo 13**

# Modelo de Regressão Múltipla

Previsão da Percentagem de Despesa Total Alocada para Alimentação num Agregado Familiar de Espanha (por mês)

**Trabalho Realizado por:**

Catarina Castanheira, n.º 92478

João Martins, n.º 93259

Joel Paula, n.º 93392

# Índice

# Introdução

O objetivo deste trabalho é a previsão da percentagem da despesa total alocada para alimentação num agregado familiar de Espanha (por mês), ou seja, implica o desenvolvimento de um modelo de regressão linear múltipla.

O *dataset* utilizado neste trabalho, *BudgetFood*, contém um conjunto de 23972 observações em 6 variáveis:

- ***Wfood:*** percentagem da despesa total que uma família gasta para alimentação /por mês
- ***Totexp:*** despesas totais do agregado familiar
- ***Age:*** idade da pessoa de referência do agregado familiar
- ***Size:*** número de elementos do agregado familiar
- ***Town:*** tamanho da cidade onde a família tem o domicílio (5 categorias: 1 - cidade pequena, … ,5 - cidade grande)
- ***Sex:*** sexo da pessoa de referência do agregado familiar (*man*, *woman*)

Neste projeto será utilizada uma subamostra deste *dataset*, composta por 1300 observações.

# Apuração do Melhor Modelo

## Análise exploratória das variáveis

### Variáveis explicativas e variável de resposta

Para o desenvolvimento do modelo de regressão múltipla definimos como variável de resposta a *wfood* – a variável dependente do nosso estudo –, e como variáveis explicativas (independentes): *totexp*, *age*, *size*, *town* e *sex*.

### Definição da amostra

Neste projeto iremos considerar somente as linhas 15601 a 16900 para a nossa amostra (amostra_grupo_13), que contém 1300 registos.

Nesta amostra a variável *wfood* apresenta um mínimo de 0 (0% da despesa total que uma família gasta para alimentação por mês) e máximo de 0.9193 (91.93% da despesa total), média de 0.4001 (40.01% da despesa total) e mediana de 0.3914 (39.14% da despesa total). Analisando os quantis, observa-se que 25% dos inquiridos refere gastar 27.29% da despesa total para alimentação por mês (0.2729) e 75% dos inquiridos referem gastar até 51.53% da despesa total (0.5153).

Ao estudar a variável *totexp* é possível verificar que o mínimo das despesas totais dos agregados familiares dos inquiridos desta amostra é 243.4€ e o máximo é 40420.6€, a média é de 5514.1€ e a mediana é 4717.2€. Analisando os quantis, observa-se que 25% dos inquiridos tem 2848.9€ de despesa total do agregado familiar e 75% dos inquiridos tem despesas até 7093€.

Quanto à variável *age*, o menor valor é 18 e o maior é 99, com uma média de 50.22 anos e mediana de 49 anos. Aqui, 25% dos inquiridos desta amostra têm 39 anos e 75% dos inquiridos têm até 60 anos.

Relativamente à variável *size*, o menor tamanho registado num agregado familiar é 1 pessoa e o maior é 14 pessoas; temos uma média de 4.16 e uma mediana de 4, sendo que 25% dos inquiridos desta amostra têm um agregado familiar composto por até 2 pessoas e 75% até 5 pessoas.

Olhando agora para a variável *town*, sendo 1 uma cidade pequena e 5 uma cidade grande, o mínimo e máximo registado nesta variável é 1 e 4 respetivamente, temos uma média de 3.17 e uma mediana de 4; 25% dos inquiridos habitam cidades até dimensão 2 e 75% dos inquiridos habitam cidades até dimensão 4.

A variável *sex*, por ser uma variável nominal, levou à criação de uma variável *dummy*, à qual pudemos depois realizar a sua análise descritiva. Nesta sub-amostra temos uma distribuição de 15.2% de mulheres e 84.8% de homens.

É também de mencionar que após a sua análise, na amostra em estudo não se observam *missing values* nem valores invulgares.

## Análise das correlações entre as variáveis

Quanto ao estudo das correlações entre a variável de resposta e as variáveis explicativas, a variável *totexp* é a que parece ter uma maior correlação (negativa) com a *wfood* (-0.49358955) o que faz sentido, uma vez que *wfood* é calculada a partir de *totexp* – é uma proporção desta. É seguida pela *age* e *town,* que têm uma pequena correlação com *wfood* (0.252 e -0.182 respetivamente).

Ao observarmos as correlações das variáveis explicativas entre si temos correlações médias (com coeficiente de correlação de *Pearson* de 0.30 a 0.49) entre as variáveis: *size* e *totexp* (0.347), *size* e *age* (-0.314) e *sex* e

*size* (-0.353). As restantes são correlações pequenas (com coeficiente de correlação de *Pearson* menor que 0.29).

Matriz de correlação de Pearson:

| | ***wfood*** | ***totexp*** | ***age*** | ***size*** | ***town*** | ***sex*** |
|---|---|---|---|---|---|---|
| ***wfood*** | 1.00000000 | -0.4935896 | 0.2522291 | 0.03367267 | -0.182467864 | 0.049767720 |
| ***totexp*** | -0.49358955 | 1.00000000 | -0.2363558 | 0.34660298 | 0.156399779 | -0.212813376 |
| ***age*** | 0.25222908 | -0.2363558 | 1.00000000 | -0.31388915 | -0.174618621 | 0.286709376 |
| ***size*** | 0.03367267 | 0.3466030 | -0.3138892 | 1.00000000 | 0.063189718 | -0.353215630 |
| ***town*** | -0.18246786 | 0.1563998 | -0.1746186 | 0.06318972 | 1.00000000 | 0.001673588 |
| ***sex*** | 0.04976772 | -0.2128134 | 0.2867094 | -0.35321563 | 0.001673588 | 1.000000000 |

## Definição do modelo de regressão múltipla

Após a análise exploratória das variáveis passámos para a escolha do modelo de regressão com melhor ajuste com base no *p-value*. Começamos por criar o modelo linear usando o *wfood* como variável de resposta e todas as restantes variáveis como variáveis explicativas, como referido acima. Observámos que todas as variáveis eram significativas, excepto a variável *sex.* Obtivemos um $R^2$ de 0.3363 e um $R^2$ ajustado de 0.3389 para este primeiro modelo (*wfood = totexp + town + sex + age + size*). Ao fazermos os *plots* dos resíduos, constatámos que estes se encontravam distribuídos em "forma de funil".

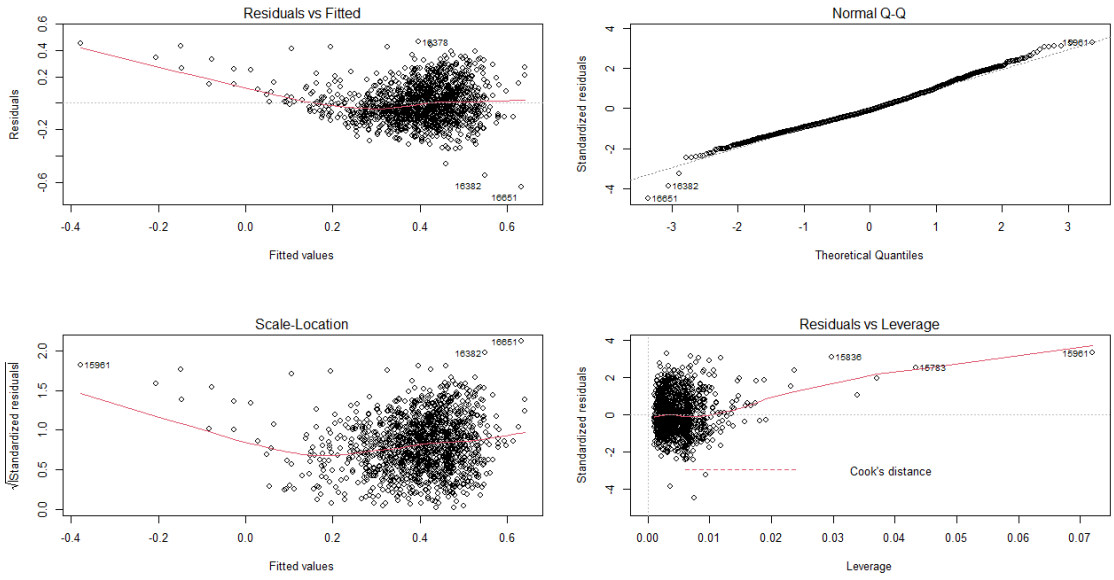


*Figura 1 – Regressão Linear*

Posto isto, construímos um modelo de regressão a partir do modelo anterior usando a regressão *stepwise* e verificámos mais uma vez que todas as variáveis, exceto a *sex,* eram significativas, porque afetavam o $R^2$ ajustado e o AIC (*Akaike Information Criterion*).

## Metodologia

Para verificarmos os pressupostos dos resíduos realizámos a média dos resíduos do modelo, os testes de *Breusch-Pagan,* de *Breusch-Godfrey* e de *Jarque Bera*.

Relativamente à média dos resíduos, obtivemos uma média nula, o que conduz à verificação do primeiro pressuposto dos modelos de regressão linear.

No teste de *Breusch-Pagan* (utilizado para verificar a homocedasticidade) a não-rejeição da $H_0$ (isto é, *p-value* > 0.05) conduz a verificação do pressuposto ($H_0$ - variância é constante). Neste caso foi rejeitado $H_0$ (*p-value* = 4.554e-11), logo este pressuposto não está a ser verificado, o que quer dizer que os resíduos são não-homocedásticos.

No teste de *Breusch-Godfrey* (usado para verificar a independência dos resíduos) a não-rejeição da $H_0$ (isto é, *p-value* > 0.05) conduz a verificação do pressuposto ($H_0$ – resíduos são independentes). Neste caso foi rejeitado $H_0$ (*p-value* = 4.391e-09), logo este pressuposto não está a ser verificado, o que quer dizer que os resíduos são correlacionados.

No teste de *Jarque Bera* (usado para verificar a normalidade dos resíduos) a não-rejeição da $H_0$ (isto é, *p-value* > 0.05) conduz a verificação do pressuposto ($H_0$ – resíduos seguem uma distribuição normal); neste caso é rejeitado $H_0$ (*p-value* = 6.158e-06), logo este pressuposto não está a ser verificado, o que quer dizer que os resíduos não seguem uma distribuição normal.

Destes testes apenas 1 dos pressupostos é cumprido, o que nos leva a concluir que o modelo não é robusto, já que é pouco eficiente e os seus intervalos de confiança não são fidedignos.

## Análise da multicolinearidade

Recorremos à função VIF (*Variance Inflation Factor*) para avaliar a multicolinearidade do nosso modelo de regressão múltipla e verificámos que não existiam variáveis multicolineares porque o VIF de todas era inferior a 5.

VIF(fit):

| totexp | age | size | town | sex |
|---|---|---|---|---|
| 1.186822 | 1.202534 | 1.301932 | 1.052620 | 1.202158 |

## Análise dos pressupostos sobre os resíduos

Os pressupostos que foram verificados para validar os modelos de regressão linear foram:

- $\mathrm{E}(\varepsilon_t) = 0$ – A média dos resíduos é nula
- $\mathrm{Var}(\varepsilon_t) = \sigma^2$ – Homocedasticidade dos resíduos - a variância dos erros é constante e finita
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ – Os resíduos são linearmente independentes
- $\varepsilon_t \sim N(0, \sigma^2)$ – Os resíduos são normalmente distribuídos

Em todos os modelos experimentados, apenas se confirmou o primeiro pressuposto.

Não foi feito nenhum teste específico para verificar a independência entre os resíduos e as variáveis independentes ($\mathrm{Cov}(\varepsilon_t, x_t) = 0$).

### Resíduos com média zero

Em qualquer um dos modelos a média dos resíduos sempre foi muito próxima de zero (a menos de uma décima). Estão bem distribuídos entre ambas as metades do gráfico:

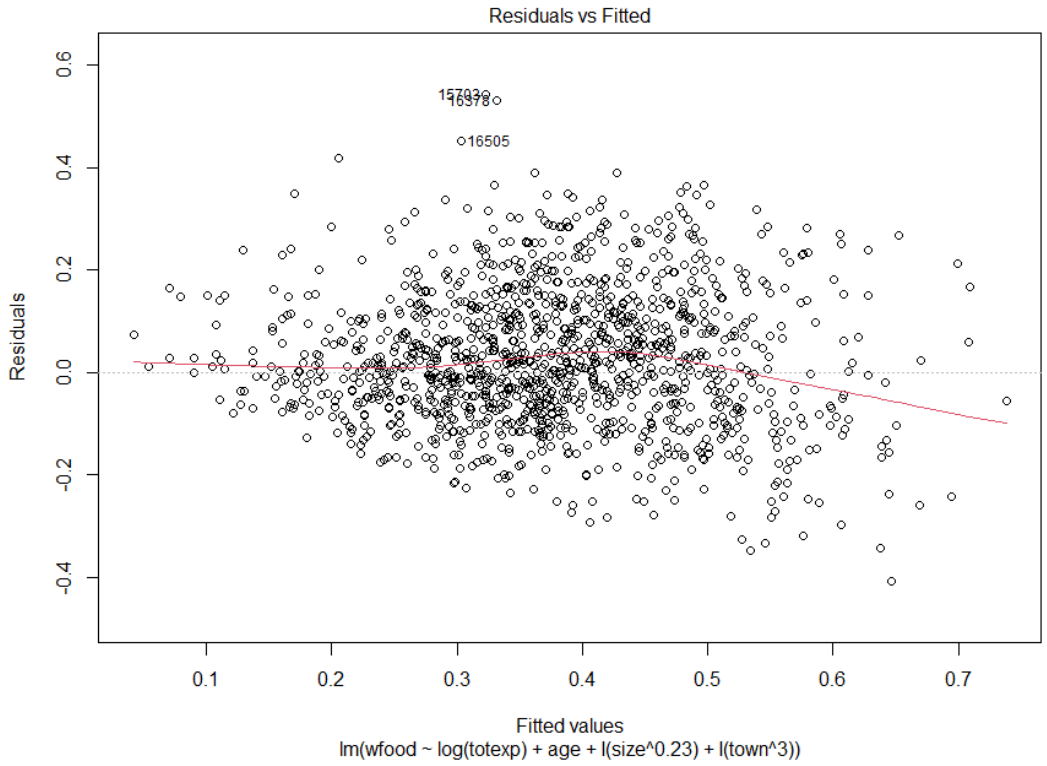


*Figura 2 - Distribuição dos resíduos vs valores preditos, com média nula*

Isso foi confirmado não só através do cálculo da média dos resíduos, mas também pelo facto de o termo constante (*intercetpt*) ser sempre não nulo ($\beta_0 \neq 0$).

## Homocedasticidade

Para testar a homocedasticidade, usámos o Teste LM de *Breusch-Pagan*. Neste teste a hipótese nula ($H_0$) é os erros serem homocedásticos. Em todos os modelos, o *p-value* deste teste foi bastante inferior a 0.05, o que

indica uma rejeição da hipótese nula e indica os resíduos como heterocedásticos. Isso era visível nos gráficos, pois os resíduos sempre apresentaram a característica forma de funil:

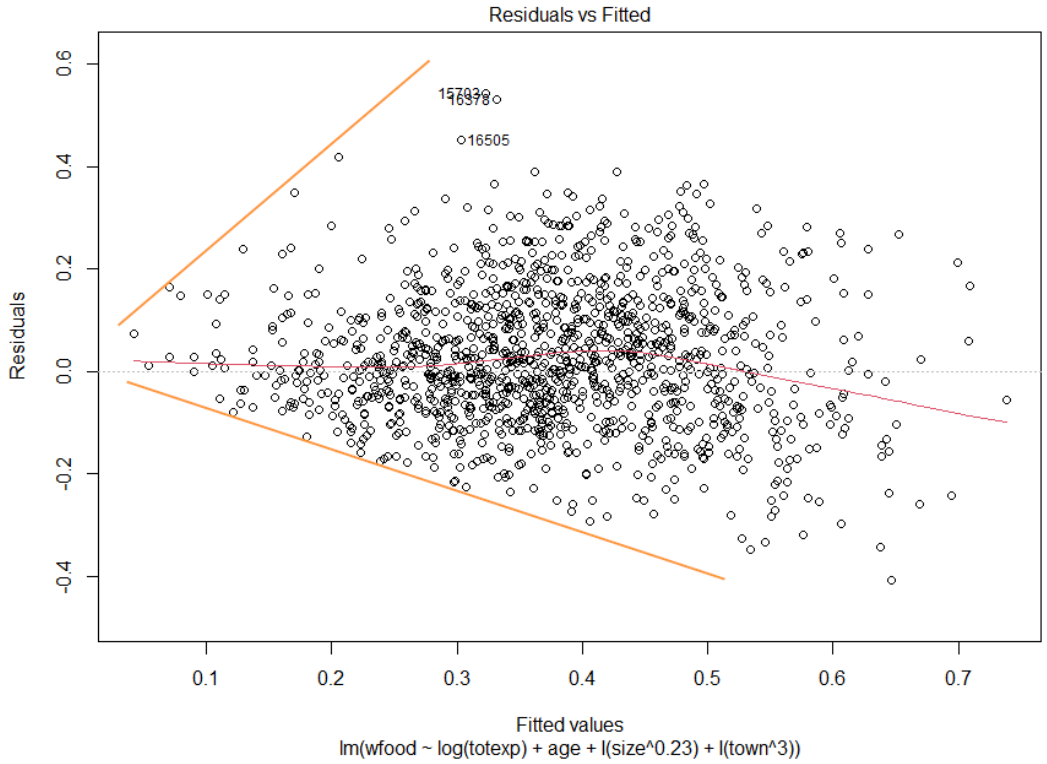


*Figura 3 - resíduos apresentam padrão de "funil"*

A consequência desta falta de homocedasticidade é que os estimadores dos modelos são pouco eficientes e os intervalos de confiança não são fidedignos.

## Independência dos resíduos

Os erros não apresentam um padrão de "ruído branco", apresentado antes o padrão de "funil". Isso é confirmado pelo teste de *Breusch-Godfrey* cuja hipótese nula ($H_0$) é de que o erro atual não é correlacionado com nenhum dos seus valores anteriores ( $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall\, j < i$). Em todos os modelos a hipótese nula sempre foi rejeitada por um *p-value* inferior a 0,05, indicando que os resíduos são correlacionados.

A consequência deste problema do modelo é que, mais uma vez, os estimadores dos modelos são pouco eficientes e os intervalos de confiança não são fidedignos.

## Resíduos normalmente distribuídos

Para fazer esta avaliação graficamente foi usado um *QQPlot*, que sempre mostrou um afastamento da curva normal:

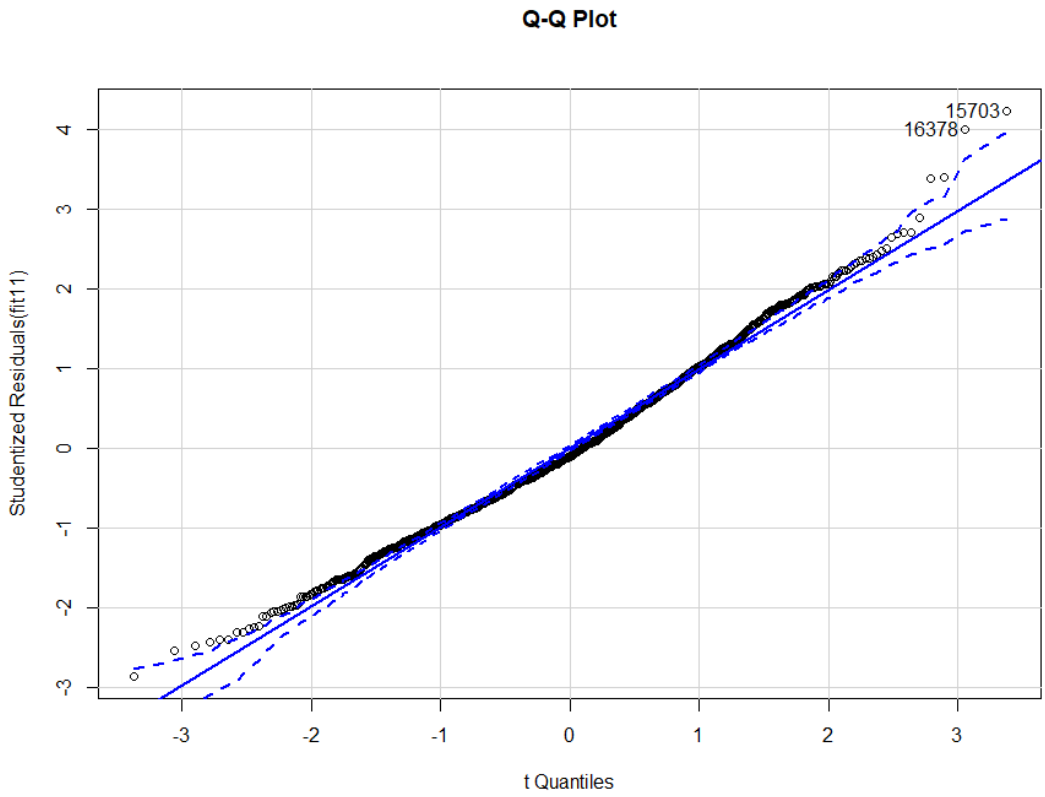


*Figura 4 - QQPlot permite verificar se os resíduos seguem uma distribuição normal e assinala os outliers*

Adicionalmente, foi executado o teste de Jarque-Bera, cuja hipótese Nula ($H_0$) é de que os resíduos seguem uma distribuição Normal. Em todos os modelos que foram testados, essa hipótese sempre foi rejeitada por *p-values* muito próximos de zero.

## Medidas corretivas

Quando os pressupostos não se verificaram, tomaram-se várias medidas, começando-se por eliminar *outliers* de grande impacto, tendo o cuidado de não eliminar demasiados registos e com isso adulterar a amostra.

### *Outliers* e influenciadores

A abordagem começou por ser o levantamento e eliminação de *outliers*. Para isso foi usado o teste de *Bonferroni*. Os primeiros *outliers* diziam respeito a famílias que não apresentavam qualquer despesa com alimentação. Seja por erro ou por outras razões, as 4 amostras que demonstravam esse comportamento foram removidas da análise, bem como dos *datasets* de treino e de teste.

Uma análise de *outliers* nas variáveis independentes com base no *boxplot* (amostras cujos valores se encontram abaixo de $q25 - 1.5\,x\,IQR$ ou acima do $q75 + 1.5\,x\,IQR$), não revelou mais valores que pudessem ser considerados erros ou valores anormais.

Uma análise à variável dependente também revelou alguns valores bem próximos de zero, sendo que uma amostra com cerca de 0.5% de despesas alimentares também foi removida das análises.

Numa segunda abordagem foram verificados os maiores influenciadores com maior distância de Cook, usando a função `influencePlot()`, e "*Hat Plots*".

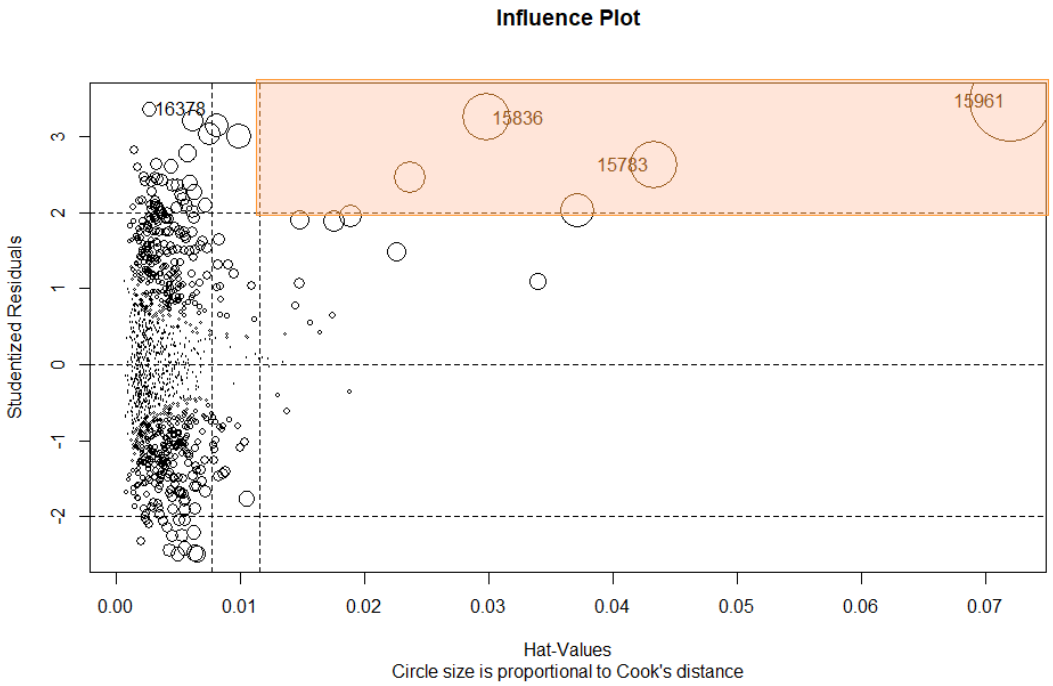


*Figura 5 - Gráfico de influência assinalando zona onde encontramos amostras outliers influenciadoras*

No caso dos gráficos de influência, foram analisadas as observações indicadas nos quadrantes superior e inferior direito. Genericamente, com distâncias de *Cook* altas em relação à média, *hat-values* superiores a 3 vezes a média e resíduos normalizados superiores a +/- 2.

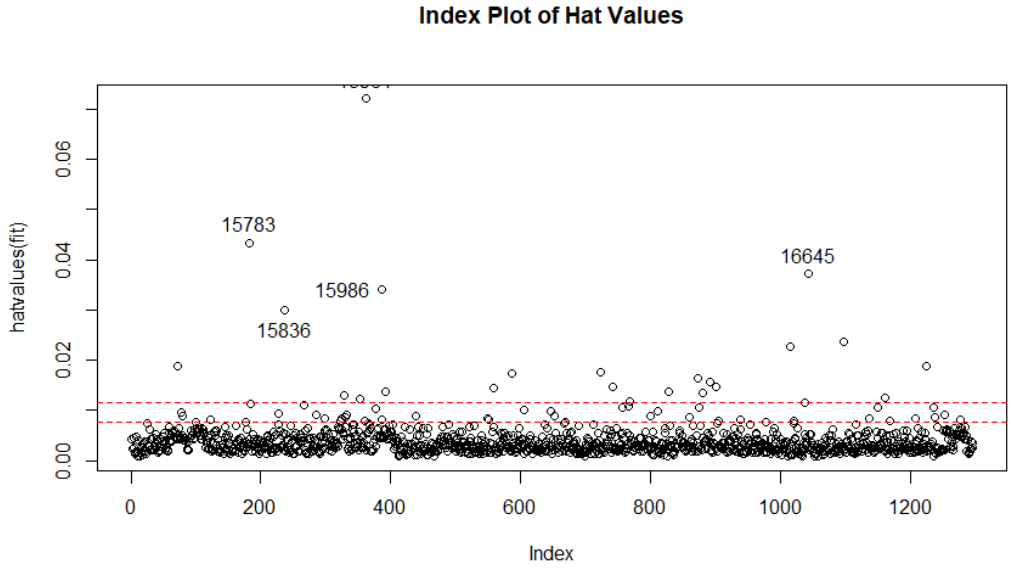


*Figura 6 - "Hat-Plot" para analisar amostras influenciadoras*

Para o caso dos *Hat-Value Plot*, o critério foi congruente – de observar as amostras cujo *hat-value* era superior a 3 vezes a média.

Foram executadas duas análises por distância de *Cook*. Uma para os casos em que a distância era superior a $4/(n - k - 1)$, em que $n$ é o tamanho da amostra e $k$ é o número de variáveis preditoras. Outra para os casos em que a distância era superior a 4 vezes a média. O número de observações abrangidas era bastante alto (73 e 42, respetivamente) e não revelavam nenhum tipo especial de padrão. Foi decidido que não seriam eliminadas estas observações por esta via, para não enviesar os resultados da futura previsão.

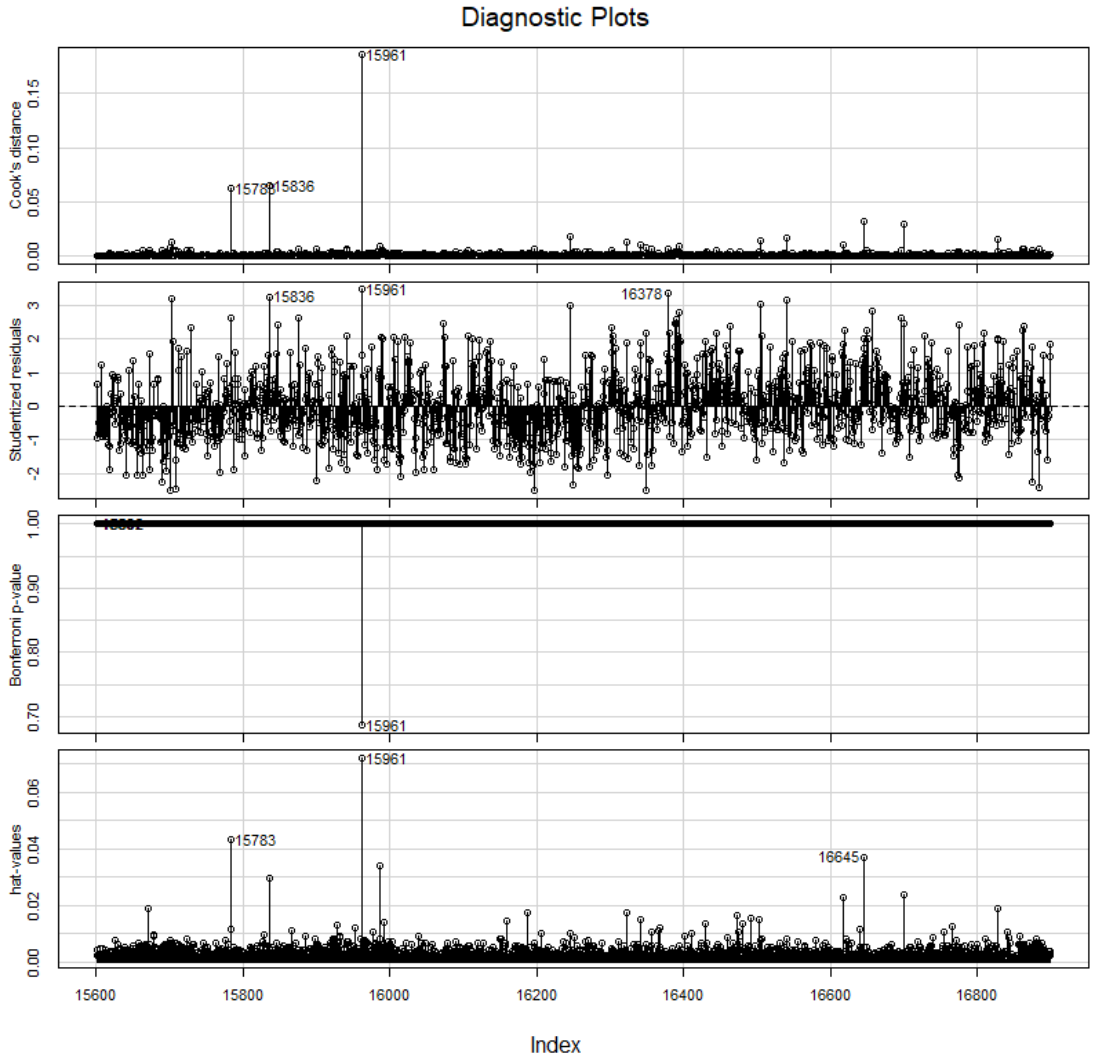


*Figura 7 - Influence Index Plot, para análise de amostras influenciadoras*

Para cada modelo, foi usada uma combinação de *outliers* que eram influenciadores, como critério de seleção das observações a eliminar.

## Transformação de Variáveis

Como medida inicial de tentativa de obter homocedasticidade, decidiu-se transformar a variável dependente (*wfood*) com a função *logit* ( $ln(Y/1 - Y)$ ). Uma vez que não resultou, experimentou-se logaritmizar a mesma.

Numa tentativa de normalizar o modelo, decidiu-se transformar também as variáveis independentes. Para isso foi usada a função `powerTransform()`, da biblioteca *cars*, e foram obtidos os $\lambda$ que permitiram determinar as transformações a aplicar a cada variável independente, segundo a tabela:

| $\lambda$ | -2 | -1 | -0,5 | 0 | 0,5 | 1 | 2 |
|---|---|---|---|---|---|---|---|
| transformação | $1/Y^2$ | $1/Y$ | $1/\sqrt{Y}$ | $\log(Y)$ | $\sqrt{Y}$ | Nada | $Y^2$ |

### Adição e remoção de variáveis

A variável independente "sex", por não ser considerada significativa nos modelos iniciais nem em diversas explorações, foi removida.

Numa tentativa de solucionar a heterocedasticidade, a não normalização dos resíduos e a autocorrelação dos resíduos, adicionaram-se algumas combinações de variáveis. Foram testadas as significâncias de todas as combinações possíveis, tendo apenas sido mantidas aquelas que eram significativas.

## Escolha do Melhor Modelo

Para analisar o potencial preditivo de cada um dos modelos criados, foi aplicada a medida de predição MAPE – *Mean Absolute Percentage Error* –, que tem por base os valores *reais* das observações na amostra (`actual`) e também dos valores previstos pela equação do modelo (`prediction`). O resultado obtido com esta métrica corresponde à proporção de erros obtidos na predição de valores para a variável de resposta *wfood*. Nesta ótica, quanto maior o valor para o MAPE, maior o erro associado ao respetivo modelo de regressão, e por isso menos útil este último será.

Além disto, teve-se em conta o valor de AIC (*Akaike Information Criterion*) para cada um dos modelos. Este estimador avalia a qualidade de cada um dos modelos, relativamente aos restantes. O menor resultado obtido significará que o dado modelo terá melhor qualidade que os restantes.

### Comparação dos modelos

Os resultados dos valores de MAPE e AIC obtidos para cada um dos modelos foi o seguinte:

| Modelo de Regressão | *Pseudo-Equação* do Modelo | MAPE | AIC |
|---|---|---|---|
| **fit** <- lm(wfood ~., data = amostra_grupo13_n) | $wfood = totexp + age + size + town + sex$ | 0.5099 | -1383.561 |
| **fit2** <- lm(wfood ~ totexp+age+size+town, data = amostra_grupo13_n) | $wfood = totexp + age + size + town$ | 0.5156 | -1427.500 |
| **fit3** <- lm(wfood ~ totexp+age+size+town, data = amostra_grupo13_n) | $wfood = totexp + age + size + town$ | 0.5079 | -1437.448 |
| **fit4** <- lm(log(wfood) ~ totexp+age+size+town, data = amostra_grupo13_n) | $\log(wfood) = totexp + age + size + town$ | 0.4168 | 1224.921 |
| **fit5** <- lm(log(wfood) ~ sqrt(totexp)+age+sqrt(size)+I (town^3), data = amostra_grupo13_n) | $\log(wfood) = \sqrt{totexp} + age + \sqrt{size} + town^3$ | 0.4167 | 1152.133 |
| **fit6** <- lm(wfood ~ totexp+totexp*size+totexp*age +totexp*town+size*age+size*to wn+age*town+age+size+town, data = amostra_grupo13_n) | $wfood = totexp + totexp \times size + totexp \times age + totexp \times town + size \times age + size \times town + age \times town + age + size + town$ | 0.4536 | -1464.144 |
| **fit7** <- lm(wfood ~ totexp+totexp*size+totexp*tow | $wfood = totexp + totexp \times size + totexp \times town + size \times age + age + size + town$ | 0.4536 | -1467.323 |

| | | | |
|---|---|---|---|
| n+size*age+age+size+town, data = amostra_grupo13_n) | | | |
| **fit8** <- lm(wfood ~ log(totexp)+totexp*size+totexp*town+size*age+age+sqrt(size)+I(town^3), data = amostra_grupo13_n) | $wfood = \log(totexp) + totexp \times size + totexp \times town + size \times age + age + \sqrt{size} + town^3$ | 0.4382 | -1536.861 |
| **fit9** <- lm(wfood ~ log(totexp)+totexp*town+age+sqrt(size)+I(town^3), data = amostra_grupo13_n) | $wfood = \log(totexp) + totexp \times town + age + \sqrt{size} + town^3$ | 0.4383 | -1540.381 |
| **fit10** <- lm(wfood ~ log(totexp)+totexp*town+age+sqrt(size)+I(town^3), data = amostra_grupo13_n, weights = 1/(sqrt(1:N))) | $wfood = \log(totexp) + totexp \times town + age + \sqrt{size} + town^3$ | 0.4217 | -1378.990 |
| **fit11** <- lm(wfood ~ log(totexp)+age+I(size^0.23)+I(town^3), data = amostra_grupo13_n, weights = w) | $wfood = \log(totexp) + age + size^{0.23} + town^3$ | 0.4522 | -1570.945 |
| **fit12** <- rlm(wfood ~ log(totexp)+age+I(size^0.23)+I(town^3), data = amostra_grupo13_n, weights = w) | $wfood = \log(totexp) + age + size^{0.23} + town^3$ | 0.4476 | -1567.547 |
| **fit13** <- lm(wfood ~ log(totexp)+age+I(size^0.23)+I(town^3), data = amostra_grupo13_n, weights = 1/h) | $wfood = \log(totexp) + age + size^{0.23} + town^3$ | 0.4366 | -2592.671 |
| **fit14** <- lm(wfood ~ log(totexp)+age+I(size^0.23)+I(town^3), data = amostra_grupo13_n, weights = 1/varfunc1^0.5) | $wfood = \log(totexp) + age + size^{0.23} + town^3$ | 0.3968 | -1515.092 |
| **fit15** <- lm(wfood ~ log(totexp)+age+I(size^0.23)+I(town^3), data = amostra_grupo13_n, weights = 1/varfunc1^0.5) | $wfood = \log(totexp) + age + size^{0.23} + town^3$ | 0.4017 | -1594.640 |

## Justificação da escolha

Através da análise dos quadros acima representados, torna-se evidente que nenhum dos modelos de regressão linear tem uma boa eficácia na predição de valores para a variável *wfood*. Olhando para a medida de MAPE, os melhores resultados implicam erros de 0.3968, 0.4017 e 0.4167 (obtidos pelos modelos fit14, fit15 e fit5, respetivamente). Por outro lado, se considerarmos os resultados do AIC, os modelos com melhor qualidade serão os fit13, fit15 e fit11. Não é claro dentro destas hipóteses de modelo, aquele que será o mais indicado.

Para complementar a decisão, foi feita uma nova análise aos valores do $R^2$ ajustado, medida que informa sobre a proporção de variação em *wfood* que é explicada por uma variação nas variáveis independentes. Neste sentido, valores mais baixos para esta medida implicam uma maior intervenção de "ruído" (erro) na determinação do valor da variável de resposta. Tendo em conta somente este prisma, os melhores modelos são os fit5 (0.4419), fit11 (0.4330), fit14 (0.4205), fit13 (0.4194).

Com isto, conclui-se que não existe à partida um melhor modelo para conduzir a previsão desejada. Iremos considerar os modelos 11, 13, 14 e 15 para conduzir a previsão da variável de resposta *wfood*.

## Treino dos Modelos e Respetivas Previsões

Para o treino dos modelos foi decidida uma partição da amostra numa razão 30/70, obtendo assim uma subamostra para treino composta por 742 observações e uma subamostra de teste composta por 554 observações, totalizando 1296.

### Resultados das Previsões

Começámos por implementar o modelo 11 na previsão da percentagem de despesa total alocada à alimentação, tendo obtido um erro nas previsões perto de 41% (MAPE = 0.4091).

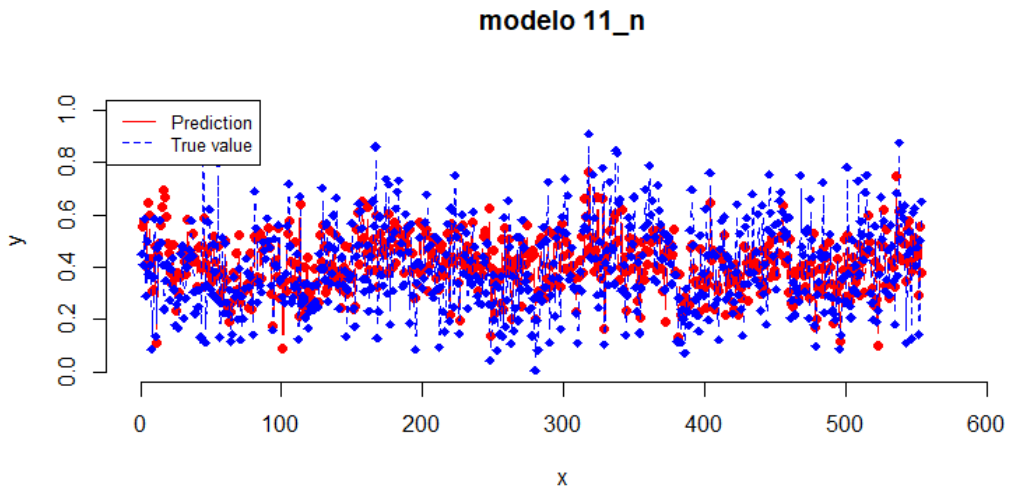


*Figura 8 - Modelo 11: Diferença entre valores reais e valores previstos para wfood*

Seguiu-se a utilização do Modelo 13, com uma qualidade na previsão muito semelhante ao anterior (MAPE = 0.4106).

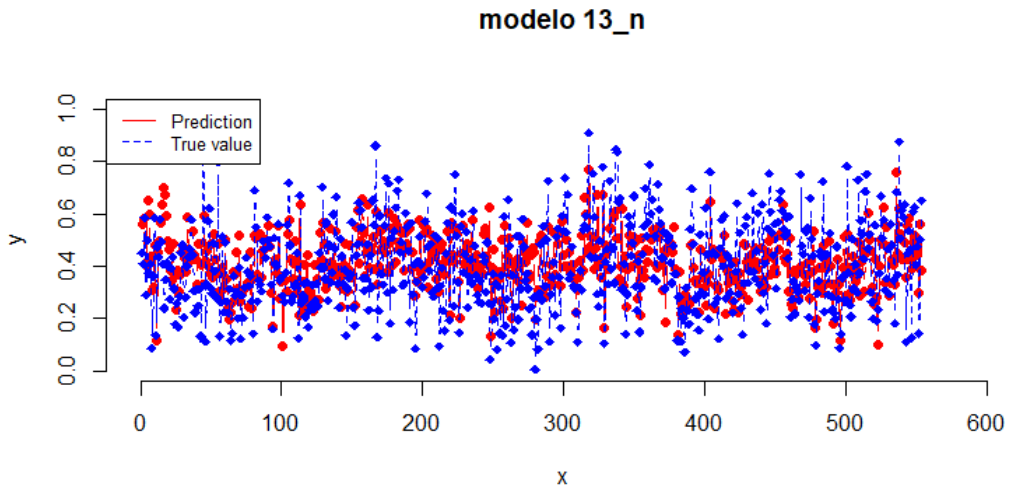


*Figura 9 - Modelo 13: Diferença entre valores reais e valores previstos para wfood*

Aplicando o Modelo 14 na previsão dos valores de *wfood*, obtemos uma ligeira melhoria na qualidade dos resultados, com um erro de previsão na ordem dos 38% (MAPE = 0.3782).

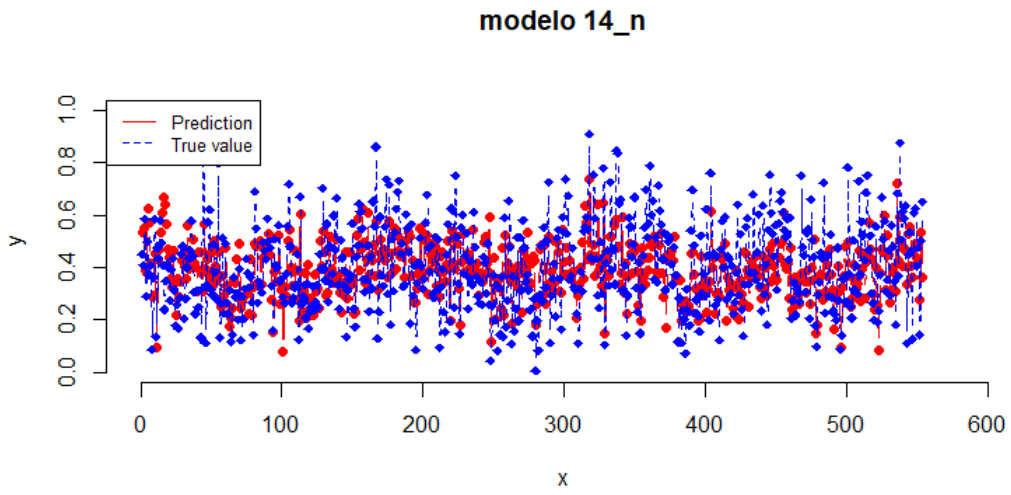


*Figura 10 - Modelo 14: Diferença entre valores reais e valores previstos para wfood*

Finalmente, aplicando o Modelo 15 na previsão, concluímos que não existe diferença na qualidade uma vez que obtemos um valor para o MAPE exatamente igual ao obtido da aplicação do modelo anterior. A remoção de influenciadores do Modelo 14 não se refletiu numa melhoria da capacidade de previsão deste Modelo 15.

# Conclusões

A criação de modelos de regressão linear e a sua aplicação na previsão em contextos reais de uma forma eficaz está sujeito à qualidade dos dados da amostra obtida (no caso deste trabalho assume-se esta qualidade) e também à verificação de pressupostos em relação ao comportamento das variáveis e dos resíduos. Uma boa capacidade na predição advém do uso de dados que representam bem a população de onde são retirados e também da nossa própria capacidade em aferir o melhor modelo representativo.

Neste projeto, apesar das medidas corretivas implementadas, não foi possível a obtenção de um bom modelo preditivo para o objetivo proposto – o modelo com o melhor desempenho aparente foi o modelo 14, com um erro em 38% dos valores. Dizemos "aparente" porque não nos podemos esquecer que existem pressupostos dos modelos de regressão linear que não foram verificados e que têm impacto na previsão em contextos reais, nomeadamente para outras amostras da mesma população. Sem os pressupostos verificados não é possível a obtenção de um modelo válido. Os pressupostos não verificados foram os da homocedasticidade dos resíduos, da independência dos resíduos e da normalidade dos resíduos.

Aliada à não verificação de todos os pressupostos, na maioria dos modelos observou-se uma predominância de ruído quando determinado o $R^2$ ajustado. Somente no modelo 12 foi conseguido um valor razoável de 0.8375; nos restantes, não se ultrapassou o valor de 0.4419 (Modelo 5) para esta medida. Contudo, não se verificando os pressupostos acima mencionados, estes valores de $R^2$ ajustados não são fidedignos.

A previsão *in-sample* e *out-of-sample* acabam por ter algumas variações, mas correspondem a ordens de grandeza iguais (o valor de MAPE varia no máximo cerca de 6 pontos percentuais para os modelos em que foi aplicada a previsão).

Tendo em conta os problemas encontrados, parece ser difícil conseguir encontrar um modelo de regressão linear que utilizando apenas as variáveis independentes disponíveis consiga explicar adequadamente a proporção de gastos em alimentação num agregado familiar de Espanha.

# Bibliografia

Kabacoff, R. I. (2015). *R in Action* (Second ed.). Shelter Island: Manning.