UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Análise da presença de variáveis mediadoras

Catarina Pereira Rodrigues da Venda

Mestrado em Estatística e Investigação Operacional
Especialização em Estatística

Dissertação orientada por:
Professora Doutora Patrícia Cortés de Zea Bermudez
Professora Doutora Luzia Augusta Pires Gonçalves

2017

Agradecimentos

Em primeiro lugar, quero agradecer aos meus pais. É a eles que devo o meu percurso até aqui. Um "obrigado" nunca será suficiente para agradecer toda a confiança que depositaram em mim, todo o esforço e investimento que fizeram e apoio incondicional que me deram. Ver a alegria e o orgulho deles com o culminar desta etapa faz-me sentir que tudo valeu (tanto!) a pena.

Às minhas orientadoras, Professora Doutora Patrícia Bermudez e Professora Doutora Luzia Gonçalves. Não podia ter escolhido melhores pessoas para me orientar nesta dissertação. Agradeço muito a amizade, a disponibilidade, a ajuda prestada, os conhecimentos partilhados e o profissionalismo. Foram incansáveis em qualquer momento, bom ou menos bom.

À minha irmã e ao meu irmão, por todo o apoio, preocupação e brincadeira. Por sentir que a minha conquista, é uma conquista para eles. Obrigada manos.

Ao Cristiano, pela paciência, compreensão e motivação. Por acreditar, até ao fim, que conseguia. Não podia ter melhor companheiro para a vida.

Ao resto da minha família, aos meus amigos (em especial, às minhas Ana, Lídia e Sandrina) e às pessoas que se cruzaram no meu percurso universitário.

Obrigada!

Catarina

Resumo

O presente trabalho centra-se no conceito de mediação, ou seja, como a relação entre duas variáveis pode ser explicada por uma ou várias variáveis, as quais se designam por variáveis mediadoras. Assim, depois de definir este conceito pormenorizadamente, é feita a distinção de outros conceitos que têm igualmente na sua base a introdução de uma ou várias "terceiras" variáveis, na relação entre duas variáveis. Aqui, apenas é desenvolvido o modelo de mediação simples, isto é, o modelo que envolve apenas uma variável mediadora, apesar da referência breve ao modelo de mediação múltipla. Os conceitos fundamentais de efeito directo, efeito indirecto ou de mediação e efeito total são também descritos. De seguida, é feito um levantamento das abordagens existentes referentes a este conceito central, havendo uma distinção entre as abordagens tradicionais - que se baseiam nas estimativas obtidas através de um conjunto de equações de regressão linear, o qual traduz o modelo de mediação simples – e a abordagem contrafactual - que se baseia no conceito de resultados potenciais. A abordagem contrafactual é mais fácil de ser aplicada a qualquer modelo estatístico.

A aplicação deste assunto é realizada com base nos dados de um inquérito aplicado a três zonas com diferentes características, da cidade da Praia, Cabo Verde, tendo como objectivo analisar as desigualdades em saúde conforme cada zona. Algumas variáveis do questionário foram escolhidas para o estudo da mediação. Assim, são analisadas três situações que têm em comum a variável independente e dependente e variam quanto à variável mediadora. Todas estas variáveis foram tratadas como binárias. O objectivo é estimar, com base nos dados, os efeitos de interesse, para que se possa concluir quanto à presença ou ausência de mediação. Essa estimação é realizada com recurso ao pacote "*mediation*" do *R*, que usa a abordagem de resultados potenciais.

Palavras-chave: Mediação, causalidade, efeitos directos e indirectos, diferença de coeficientes e produto de coeficientes, resultados contrafactuais.

Abstract

The present work focuses on the concept of mediation, that is, how the relationship between two variables can be explained by one or more variables, which are called mediating variables. Thus, after defining this concept in more detail, a distinction of other concepts that also have at their base the introduction of one or more "third" variables in that simple relationship between two variables, is made. Here, only the simple mediation model is developed, that is, the model that involves only one mediating variable, despite the brief reference to the multiple mediation model. The fundamental concepts of direct effect, indirect or mediation effect and total effect are also introduced. Next, a survey of existing approaches to this central concept is made. On the one hand, several traditional approaches are enumerated, all based on the estimates obtained through a set of regression equations, which translate the simple mediation model; on the other, the counterfactual approach is defined, which is based on the concept of potential results, which is easier to apply to any statistical model.

The application of this subject is based on the data from a survey applied in three zones, with different characteristics, of the city of Praia, Cape Verde, in order to analyze health inequalities according to each zone. Based on the survey and in the variables withdrawn, some variables were chosen. Thus, three situations are analyzed, that has in common the independent variable and the dependent variable and varies in the mediating variable. All these variables were treated as binary. The objective is to estimate, based on the data, all the effects of interest, so that it can be concluded on the presence or absence of mediation. This estimation is done using the mediation package of R, which is based on the potential results approach.

Keywords: Mediation, causality, direct and indirect effects, difference in coefficients and product of coefficients, contrafactual outcomes.

Índice

Lista de tabelas

Lista de figuras

Lista de siglas

ACME – Average Causal Mediation Effect

ADE – Average Direct Effect

ATE – Average Treatment Effect

IC – Intervalo de Confiança

OLS - Ordinary Least Squares

SEM – Structural Equation Modeling

SUTVA – Stable Unit Treatment Value Assumption

1. Mediação

1.1. O conceito de mediação e a causalidade implícita

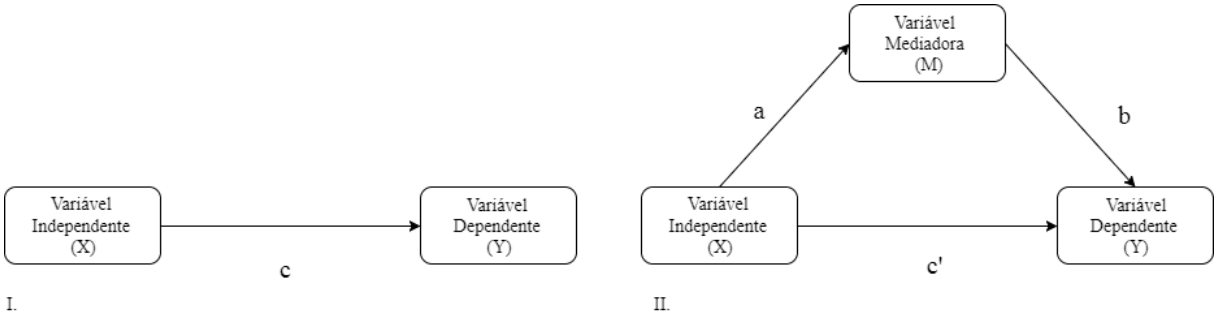
Segundo Baron e Kenny (1986: p.1167) uma variável é designada de mediador "na medida em que ela explica/é responsável pela relação entre o preditor e o critério" (Preacher e Hayes, 2004: p.717). Ou seja, o conceito de mediação refere que "uma variável independente (X) afecta uma variável dependente (Y) através de uma ou mais potenciais variáveis intervenientes, ou mediadores (M)" (Preacher e Hayes, 2008: p.879). Implícitos ao conceito de mediação estão os conceitos de causalidade e temporalidade. São vários os autores que o referem, indicando, por um lado, que a variável dependente é causada pela(s) variável(is) mediadora(s), a(s) qual(is) é(são) causada(s) pela variável independente; e por outro, que por ordem de ocorrência temporal, estão variável independente, mediador(es) e variável dependente, isto é, a variável independente precede temporalmente o(s) mediador(es), o(s) qual(is) precede(m) temporalmente a variável dependente (VanderWeele e Vansteelandt, 2009; Newsom, 2015; MacKinnon, Fairchild e Fritz, 2007; Holmbeck 1997;[1]).

Os significados de associação (ou correlação) e causalidade são diferentes: se duas variáveis estiverem relacionadas de forma causal, por exemplo X em Y, a associação que lhes está implícita, normalmente, não serve para avaliar essa relação causal, visto essa associação poder dever-se à causa inversa (efeito de Y em X) ou ao efeito de confundimento de uma outra variável (Abadie, 2005). Ou seja, quando existe uma alteração numa dessas variáveis não implica obrigatoriamente que cause uma alteração na outra variável. Este é um "aforismo fundamental em ciências sociais" (Esarey, 2015: 1). Devido a essa causalidade, está associada uma temporalidade ao conceito de mediação.

1.2. Mediação simples e mediação múltipla

O conceito de mediação remete imediatamente para uma distinção considerando o número de mediadores da relação X-Y. Caso a relação entre X e Y envolva uma variável de mediação, está-se na presença de mediação simples; caso o processo de mediação envolva mais que uma variável mediadora, está-se na presença de mediação múltipla (Preacher e Hayes, 2004; MacKinnon, 2008; Preacher e Hayes, 2008). Na aplicação do presente estudo é utilizado um único mediador, razão pela qual o modelo de mediação simples é abordado mais detalhadamente.

1.3. Modelo de mediação simples



Adaptado de Preacher e Hayes, 2008.

Figura 1.1 - Representação do modelo simples e do modelo de mediação simples. Diferença entre o efeito causal de uma variável independente (X) sobre uma variável dependente (Y), sem - situação I (modelo simples) - e com a introdução da variável mediadora (M) - situação II (modelo de mediação simples).

[1]Vide: http://www.goldsteinepi.com/blog/epivignettesmediationframeworksandanalysis e http://davidakenny.net/cm/mediate.htm.

Na figura 1.1 são denotados os efeitos intervenientes num modelo simples e de mediação simples por letras comummente usadas na área da Psicologia (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008). A situação I corresponde ao modelo simples (não mediado), no qual c representa o efeito total, isto é, o efeito causado por X em Y, ou, equivalentemente, a relação total entre somente estas duas variáveis. A situação II corresponde ao modelo de mediação simples, devido à introdução de uma variável mediadora na relação entre a variável independente e dependente. Consequentemente, o efeito total é decomposto em dois efeitos: o efeito directo (estabelece a ligação directa de X a Y) e o efeito indirecto (a ligação entre X e Y é estabelecida através da variável mediadora) - situação II da figura 1.1. O efeito directo, representado por c', considera a inclusão da variável mediadora no modelo, correspondendo à relação directa entre a variável independente e dependente, mantendo fixo M. Ou seja, é o efeito de X em Y devido a outras causas que não a mediadora. Pelo facto de considerar a mediadora, c' difere de c, pois correspondem a relações diferentes. O efeito indirecto envolve as letras a e b. Enquanto a representa o efeito da variável independente na variável mediadora, b representa o efeito da variável mediadora na variável dependente, mantendo fixa a variável independente. O efeito indirecto é o efeito de X em Y devido e explicado pela variável mediadora. Como se verá pormenorizadamente, existem duas formas de o calcular (MacKinnon, Krull e Lockwood, 2000; Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008; Newsom, 2015;[2]).

Considerando as relações subjacentes ao modelo de mediação simples esquematizado na situação II da figura 1.1, é possível traduzi-lo através do seguinte conjunto de modelos de regressão linear:

$$\mathrm{Y} = i_1 + cX + \varepsilon_1 \quad (1.1)$$

$$\mathrm{Y} = i_2 + c'X + bM + \varepsilon_2 \quad (1.2)$$

$$\mathrm{M} = i_3 + aX + \varepsilon_3 \quad (1.3).$$

Os coeficientes de cada equação, nomeadamente, a, b, c e c', representam os efeitos definidos. O modelo inclui, para além dos coeficientes de intersecção, os respectivos erros (MacKinnon, 2008; MacKinnon, Fairchild e Fritz, 2007).

O conjunto de modelos não admite interacções entre as variáveis, reflectindo apenas relações lineares entre elas: dois modelos de regressão linear simples e um modelo de regressão linear múltipla. Relativamente aos erros, pressupõe-se que não são correlacionados, seguindo uma distribuição Normal com valor esperado nulo e variância constante, σ^2 (homocedasticidade) (MacKinnon, Fairchild e Fritz, 2007; Mackinnon, 2008).

O conjunto de equações (1.1), (1.2) e (1.3) é representativo do modelo de mediação simples, assumindo vários pressupostos, entre os quais se destacam:

Pressuposto 1: O modelo está correctamente especificado se reflectir os conceitos de temporalidade e de causalidade, subjacentes à mediação. Isto é, não devem ocorrer erros de especificação relativamente à ordem temporal nem à direcção causal: a ordem deverá ser X - M – Y e o efeito causal deverá ser no sentido X → M → Y, embora possa haver uma causalidade recíproca entre M e Y (MacKinnon, Fairchild e Fritz, 2007;[2]).

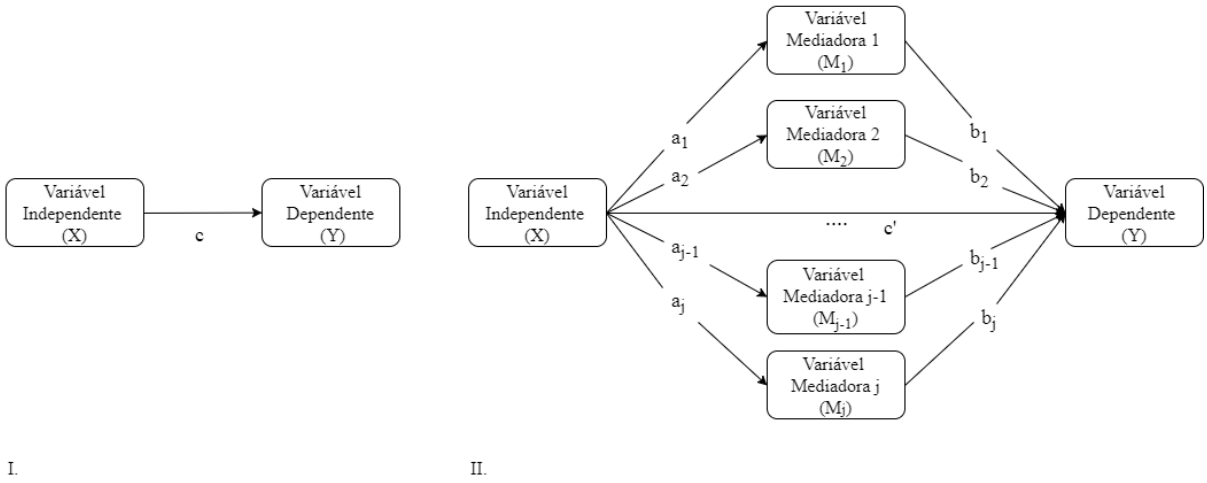
Pressuposto 2: As variáveis envolvidas, X, M e Y, devem ser mensuradas de forma válida e fiável, de modo a que as estimativas dos efeitos não sejam enviesadas (Preacher e Hayes, 2004; Mackinnon, 2008;[2]).

[2] Vide: http://davidakenny.net/cm/mediate.htm e http://davidakenny.net/cm/MediationN.ppt.

Pressuposto 3: O modelo de mediação simples considera todas as variáveis relacionadas com as três variáveis envolvidas no modelo de mediação susceptíveis de afectar as respectivas relações (X-M, M-Y e X-Y). Isto é, nenhuma variável importante poderá ser omitida, devendo ser controlada (fixa). Estas variáveis são, geralmente, designadas de factores de confundimento, os quais são mais pormenorizadamente abordados na secção 1.5.1 (Mackinnon, 2008;[2]).

A estimação dos parâmetros (efeitos) do modelo considera a distribuição das variáveis. Geralmente, e neste contexto, assume-se que qualquer das três variáveis envolvidas é contínua, seguindo uma distribuição Normal, em particular as variáveis dependente (Y) e mediadora (M) (MacKinnon, 2008). Ou seja, uma variável independente (X) com outro suporte não influencia o modelo de mediação simples definido. Nestas circunstâncias, a estimação dos efeitos pode ser realizada recorrendo ao *OLS* (método dos mínimos quadrados ordinários, do inglês *Ordinary Least Squares*), sendo aplicado outros métodos noutros contextos, como a regressão logística, a modelação de equações estruturais (*SEM*, do inglês *Structural Equation Modeling*), etc (Preacher e Hayes, 2008; MacKinnon, 2008;[2]).

1.4 Modelo de mediação múltipla



Adaptado de Preacher e Hayes, 2008.

Figura 1.2 - Representação do modelo simples e do modelo de mediação múltipla. Diferença entre o efeito causal de uma variável independente (X) sobre uma variável dependente (Y), sem - situação I (modelo simples) - e com a introdução de várias variáveis mediadoras (M_i) - situação II (modelo de mediação múltipla).

Num modelo de mediação múltipla (situação II – figura 1.2), a existência de j variáveis mediadoras da relação X-Y, gera j coeficientes a e j coeficientes b. Analogamente ao modelo de mediação simples, c e c' representam os efeitos total e directo, respectivamente, existindo também duas formas de calcular o efeito indirecto total (MacKinnon, 2008; Preacher e Hayes, 2008).

1.4.1. Exemplo de mediação múltipla

Como primeiro exemplo de mediação múltipla pode-se referir o estudo desenvolvido por Harris e Rosenthal (1985). Como variáveis independente e dependente consideraram, respectivamente, a expectativa de um professor em relação a um aluno e o sucesso escolar do aluno. Quatro variáveis foram apresentadas como mediadores: uma tendência para os professores ensinarem mais matéria, inclusive matéria mais difícil a alunos com elevada expectativa; uma tendência para dar um maior número de oportunidades de resposta a esses alunos; um *feedback* mais diferenciado do professor em relação a esses alunos e uma tendência para serem mais atenciosos com eles (MacKinnon, 2008). Outro estudo foi desenvolvido por Reynolds et al. (2004), o qual consiste em compreender o efeito do aumento do consumo de fruta e vegetais por parte das crianças na escola no consumo de fruta e vegetais pelas crianças. Como mediadores desta relação foram considerados a disponibilidade de frutas e vegetais em casa, o conhecimento da quantidade de frutas e vegetais a consumir de forma a reflectir-se na saúde e o consumo diário de fruta e vegetais por parte dos pais (Preacher e Hayes, 2008).

1.5. Diferença entre mediação e outros conceitos

O conceito de mediação remete para a introdução de uma (ou mais, no caso de mediação múltipla) terceira variável no modelo que traduz a relação entre duas variáveis. No entanto, esta definição geral está subjacente a outros conceitos, como é o caso do confundimento, da moderação e da definição de covariáveis (MacKinnon, Krull e Lockwood, 2000; MacKinnon, Fairchild e Fritz, 2007).

1.5.1. Mediador e factor de confundimento

Uma característica comum entre mediação e confundimento é a presença de causalidade. No entanto, a sua forma de manifestação nos dois conceitos é diferente (situações I e II da figura 1.3). Enquanto na mediação está implícito uma sucessão causal - X causa M, o qual causa Y-, no confundimento esta sucessão causal não se verifica (MacKinnon, Krull e Lockwood, 2000). Numa relação entre duas variáveis X-Y, uma variável é considerada factor de confundimento por se relacionar com estas duas, causando-as (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Wu, 2010). Além disso, um factor de confundimento é uma variável que confunde a relação entre as duas variáveis envolvidas no modelo, provocando uma falsa relação entre elas, podendo aumentar ou diminuir a intensidade dessa relação (MacKinnon, Krull e Lockwood, 2000; MacKinnon, 2008; Wu, 2010). Consequentemente, sempre que num modelo estiver presente um factor de confundimento que seja observável, este deve ser controlado (fixo). O grupo constituído pelos factores de confundimento impossíveis de observar ou mensurados incorrectamente constitui o confundimento residual (Wu, 2010).

MacKinnon, Krull e Lockwood (2000) e Wu (2010) referem um exemplo de comparação entre os dois conceitos, considerando um modelo constituído por duas variáveis: o salário de um indivíduo (variável independente) e a probabilidade do mesmo contrair cancro (variável dependente). Ao se introduzir a variável idade nesta relação, verifica-se que actua como factor de confundimento: por um lado, um indivíduo com mais idade, geralmente, tem maior experiência profissional, pelo que há uma tendência para que tenha um salário superior comparativamente a um indivíduo mais jovem; por outro, indivíduos mais jovens são, normalmente, menos vulneráveis a contrair cancro comparativamente a indivíduos mais idosos. Devido à sucessão causal que está subjacente à definição de mediação, é ilógico considerar a idade como variável mediadora.

1.5.2. Mediador e moderador

Ao conceito de moderação não está implícito o conceito de causalidade: um moderador, quando introduzido num modelo com duas variáveis, não é causado pela variável independente nem causa a variável dependente - situação III da figura 1.3 (MacKinnon, Fairchild e Fritz, 2007; Wu, 2010). Um efeito moderador provoca uma variação no efeito da variável independente sobre a variável dependente conforme os diversos valores que a variável introduzida assuma (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Wu, 2010). Como a variação no efeito traduz-se em termos de direcção (inversão da relação X-Y) ou de intensidade (enfraquecimento ou fortalecimento da relação X-Y) (MacKinnon, 2008; [3]), uma designação alternativa para moderador é modificador de efeito (MacKinnon, Fairchild e Fritz, 2007). A moderação é testada com a introdução, no modelo, de um termo de interacção entre a variável independente e a moderadora: se o coeficiente associado for estatisticamente significativo existe moderação[3]; caso o termo de interacção não seja estatisticamente significativo é retirado do modelo (Wu, 2010). Por este motivo a moderação é designada por interacção (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Wu, 2010). O modelo de moderação é dado por:

$$Y = \beta_0 + \beta_1 X + \beta_2 Mod + \beta_3 X.Mod + e \quad (1.4)$$

(Prado, Korelo e Silva (2014); MacKinnon, 2008; Wu, 2010; [3]).

As variáveis Y, X e Mod correspondem, respectivamente, às variáveis dependente, independente e moderadora; e representa o erro; os $\beta_i, i = 0,1,2,3$ correspondem aos coeficientes da regressão, representando as relações entre as variáveis envolventes. Por exemplo, o coeficiente β_3 representa o efeito da interacção entre as variáveis independente e moderadora sobre a variável dependente (MacKinnon, 2008; Prado, Korelo e Silva, 2014).

Um exemplo comparativo entre os conceitos de mediação e moderação é referido por Holmbeck (1997), correspondendo a um estudo de Fauber et al. (1990). As variáveis independente e dependente consistem no conflito conjugal e na forma de adaptação das crianças, respectivamente. Como variável mediadora é referida a qualidade dos progenitores enquanto pais. Conclui-se que o conflito conjugal influencia a qualidade dos cônjuges enquanto pais o que por sua vez influencia a adaptação das crianças. Como variável moderadora desta relação é referida a estrutura familiar, isto é, se na família os pais são ou não divorciados. Se a relação X-Y variar conforme os valores obtidos de uma terceira variável introduzida, os quais correspondem, neste caso, a "família com divórcio" ou a "família sem divórcio", conclui-se que a variável é um moderador. Admitindo que esta relação é apenas válida para "famílias com divórcio" então a estrutura familiar consiste num moderador da relação.

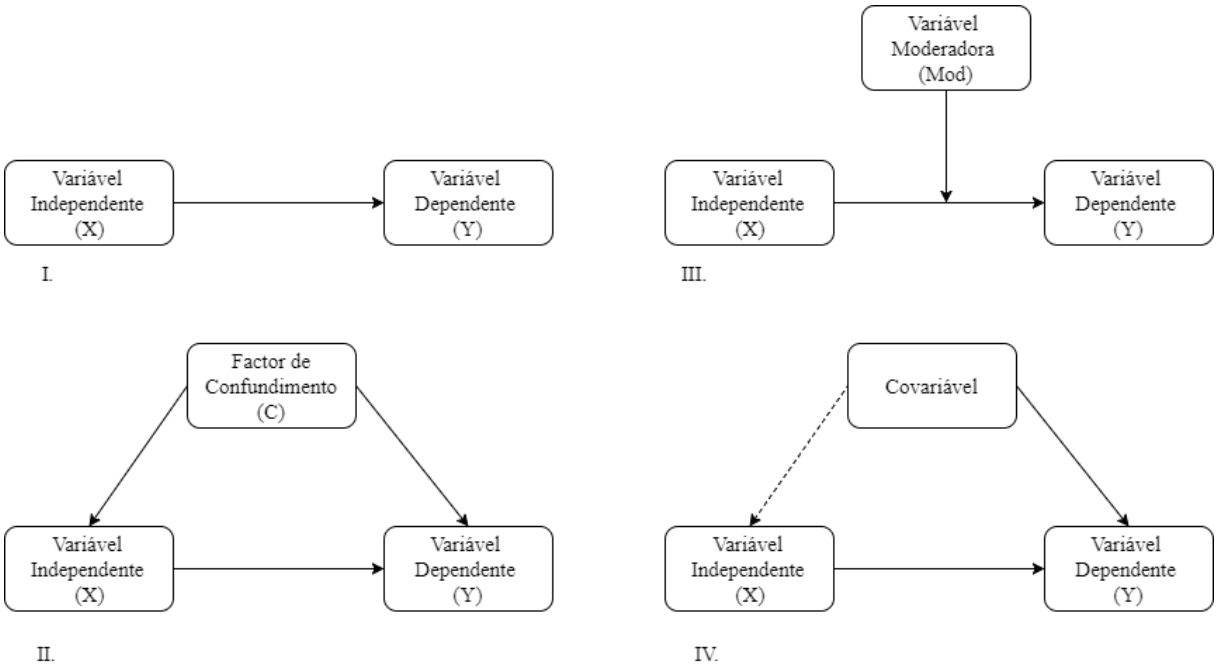
1.5.3. Mediador e covariável

As covariáveis, diferentemente do conceito de confundimento e moderação, não alteram praticamente a relação X-Y. Ao invés, permitem compreender melhor esta relação, quando introduzidas (MacKinnon, Fairchild e Fritz, 2007). Ou seja, a covariável relaciona-se com a variável dependente (Y) e, identicamente à variável independente X, prediz Y. No entanto, normalmente não está relacionada com a variável independente X, quando muito relaciona-se com X de tal forma (pouco) que não provoca uma alteração significativa na relação X-Y (Mackinnon, 2008).

As situações II e IV da figura 1.3 apresentam a diferença entre uma covariável e um factor de confundimento: uma covariável é considerada um factor de confundimento a partir do momento em

[3] Vide: https://marketinganpad.files.wordpress.com/2015/09/mediacao-e-moderacao-anpad-valter-afonso-vieira.pdf

que a relação entre a covariável e a variável independente X é suficientemente forte, de tal forma que a covariável para além de causar Y, causa X, confundindo a relação subjacente ao modelo simples X-Y (Mackinnon, 2008).



Adaptado de Wang, 2012.

Figura 1.3 - Representação do modelo simples, de confundimento, de moderação e com covariáveis. Diferença entre o efeito de uma variável independente (X) sobre uma variável dependente (Y), sem - situação I – e com a introdução de uma terceira variável que não seja mediadora - factor de confundimento (situação II), moderador (situação III) e covariável (situação IV).

2. Abordagem tradicional à análise de mediação causal

Temporalmente é possível identificar duas diferentes formas de abordar a análise de mediação causal. Tradicionalmente, esta é avaliada através do conjunto de equações lineares (1.1), (1.2) e (1.3), sendo a metodologia dos quatro passos a mais utilizada. No entanto, outras alternativas a esta abordagem padrão, com base nas mesmas equações, são mencionadas como preferíveis baseando-se, as principais, na estimação do efeito indirecto (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008; Newsom, 2015;[2]). Contemporaneamente esta é abordada considerando o conceito de resultados potenciais.

2.1. Abordagem dos passos causais ou método dos quatro passos

A abordagem mais frequentemente utilizada, e portanto a que mais se destaca de entre as referências clássicas (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008), é a metodologia que se encontra descrita no trabalho de Baron e Kenny (1986) (Preacher e Hayes, 2008; Newsom, 2015; Mackinnon, 2008). Genericamente analisam-se as equações de regressão estimadas, de forma a verificar, em cada passo, se os seus coeficientes são estatisticamente significativos (Newsom, 2015).

Os quatros passos desta abordagem consistem e são sucintamente descritos como:

1) Verificar a significância estatística do coeficiente c, através da estimação da equação (1.1).

2) Verificar a significância estatística da relação entre X e M, isto é, do coeficiente a, através da estimação da equação (1.3).

3) Verificar a significância estatística da relação entre M e Y, isto é, do coeficiente b através da estimação da equação (1.2), mantendo fixo X.
Ajustando, ao invés da equação (1.2), uma equação de regressão simples da forma:

$$Y = i_3 + bM + e_3 \quad (2.1),$$

não seria adequado. Como o objectivo deste passo é verificar se a relação entre M e Y é estatisticamente significativa, é necessário considerar na equação a variável X, fixando-a, pois as variáveis M e Y podem estar relacionadas por terem como "causa" comum a variável X (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008;[2]).

4) Verificar a significância estatística do coeficiente c' através da estimação da equação (1.2), mantendo M fixo. O objectivo é concluir sobre a existência de mediação completa ou parcial. Na mediação completa a estimativa de c' é nula[2], indicando que o efeito da variável independente na variável dependente é explicado na totalidade pela variável mediadora. Na mediação parcial a estimativa de c' é inferior à estimativa de c (em valor absoluto)[4], isto é, a relação X-Y é parcialmente explicada pela variável mediadora (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; Newsom, 2015;[5])

Esta é a versão mais utilizada do método dos quatro passos. Alguns trabalhos anteriores a esta versão, como Judd e Kenny (1981a), Judd e Kenny (1981b) ou James e Brett (1984) baseiam-se igualmente na estimação de uma série de equações de regressão, por passos, e na verificação da significância estatística dos respectivos coeficientes, como forma de avaliar a mediação. No entanto, diferenciam-se

[4] Vide: http://davidakenny.net/cm/MediationN.ppt.
[5] Vide: http://davidakenny.net/cm/mediate.htm

em certos aspectos da metodologia de Baron e Kenny (1986) (MacKinnon, Fairchild e Fritz, 2007; Mackinnon, 2008; [2]). Por exemplo, o quarto passo como foi apresentado, difere no trabalho de Judd e Kenny (1981b) no facto de exigir que c' não fosse significativo, isto é, apoia somente a mediação como forma de mediação completa (MacKinnon, 2008).

2.1.1. Mediação consistente e inconsistente

O método dos quatro passos apenas admite casos de mediação consistente, isto é, situações em que ao se introduzir, num modelo simples, uma variável mediadora, as estimativas dos efeito directo e indirecto apresentem a mesma direcção de efeito, ou seja, sinais idênticos. Caso contrário, está-se na presença de mediação inconsistente (MacKinnon, Krull e Lockwood, 2000; [2]). Na mediação inconsistente, como os sinais dos efeitos directo e indirecto são contrários, verifica-se uma tendência para os dois efeitos se compensarem. Visto a soma do efeito directo e do efeito indirecto constituir o efeito total da variável independente sobre a dependente, o efeito total virá reduzido ou mesmo nulo, ou seja, estatisticamente não significativo. Consequentemente, o primeiro passo da abordagem tradicional não é cumprido, concluindo-se, erroneamente, a ausência de mediação (MacKinnon, Fairchild e Fritz, 2007;[2]). Com isto, vários autores consideram que o primeiro passo da abordagem tradicional não é fundamental para avaliar a existência de mediação (Preacher e Hayes, 2008). Contudo, é nos modelos que envolvem vários mediadores que se regista uma maior probabilidade de ocorrência de mediação inconsistente. Nos modelos de mediação múltipla, a mediação inconsistente manifesta-se quando pelo menos um desses efeitos indirectos tenha uma direcção oposta (sinal oposto) a todos os outros efeitos (indirectos ou directo) visto existirem tantos efeitos indirectos quanto o número de mediadores envolvidos. Logo, é possível a existência de vários mediadores inconsistentes, na mediação múltipla (MacKinnon, Fairchild e Fritz, 2007).

Um exemplo considera um modelo constituído pelo stress de um indivíduo, como variável independente, a forma como este lida com o stress, como variável mediadora, e o seu humor como variável dependente. Verifica-se que o efeito directo e o efeito de mediação (ou indirecto) apresentam sinais simétricos, representando um caso de mediação inconsistente: mantendo a variável mediadora controlada, quanto maior o stress de um indivíduo, maior a sua má disposição; considerando a variável mediadora, quanto maior o stress de um indivíduo, geralmente a sua capacidade de lidar com essa situação é maior, o que reflectir-se-á de forma positiva no seu humor.[5]

2.2. Alternativa à abordagem dos passos causais

Os métodos alternativos à metodologia dos quatro passos consideram a estimação dos coeficientes das equações (1.1), (1.2) e (1.3), contudo não utilizam os critérios anteriores. Ao invés, conciliam os coeficientes apropriados de forma a estimar formas de mediação e verificar a sua significância estatística (Preacher e Hayes, 2004; Preacher e Hayes, 2008; Newsom 2015). A proporção mediada consiste numa das medidas para avaliar a mediação[5], no entanto, o efeito indirecto é a medida mais comum. Para estimar o efeito indirecto são referidos os métodos do produto de coeficientes e da diferença de coeficientes, acompanhados pelos respectivos testes de significância (MacKinnon, Fairchild e Fritz, 2007; Newsom, 2015). Outros métodos foram propostos, como o teste de significância conjunta para o efeito indirecto definido como o produto de coeficientes e o teste baseado na distribuição do produto de coeficientes.[2] No entanto, os métodos *Bootstrap* e de *Monte Carlo* são os mais utilizados (MacKinnon, Fairchild e Fritz, 2007; Preacher e Hayes, 2008; Newsom, 2015;[2]).

2.2.1. Metodologia do produto de coeficientes e respectivo teste de significância estatística

A metodologia do produto de coeficientes, no contexto do modelo de mediação simples (figura 1.1 – situação II), baseia-se na definição do efeito mediado como a multiplicação dos coeficientes de regressão a e b (MacKinnon, 2008; Newsom 2015;[5]).

De notar, que no modelo de mediação múltipla, a existência de vários mediadores provoca a existência do mesmo número de efeitos indirectos, cada um especificado conforme o mediador correspondente. Como no modelo de mediação simples o efeito indirecto pode ser traduzido como ab, no modelo de mediação múltipla o efeito indirecto específico através do mediador i pode ser definido como o produto dos coeficientes $a_i b_i, i = 1, \dots, j$. A soma de todos os efeitos indirectos específicos perfaz o efeito indirecto total (MacKinnon, 2008; Preacher e Hayes, 2008). Exemplificando através da figura 1.2 – situação II, ao supor a existência de dois mediadores na relação X-Y, o efeito indirecto através da primeira variável mediadora traduz-se como o caminho $a_1 b_1$, enquanto o efeito indirecto através da segunda variável mediadora traduz-se como o caminho $a_2 b_2$. A soma destes dois caminhos perfaz o efeito indirecto total, neste exemplo.

Retomando a mediação simples, a avaliação da significância estatística do efeito mediado pode ser efectuada por duas vias. A primeira consiste na via "tradicional", que considera a construção do respectivo intervalo de confiança para ab: se o zero estiver contido no intervalo, o efeito não é estatisticamente significativo. O IC é dado por:

$$\left(\hat{a}\hat{b} - z_{1-\frac{\alpha}{2}} * s_{\hat{a}\hat{b}}\,, \hat{a}\hat{b} + z_{1-\frac{\alpha}{2}} * s_{\hat{a}\hat{b}}\right) \quad (2.2),$$

onde α representa o nível de significância, $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição Normal reduzida e $s_{\hat{a}\hat{b}}$ o erro padrão de $\hat{a}\hat{b}$ (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008).

A segunda via foi proposta por Sobel (1982), designando-se por teste de Sobel. O objectivo do teste de Sobel é verificar se o efeito indirecto é estatisticamente significativo, assumindo-se como hipótese nula que o efeito é nulo. A avaliação da significância estatística é possível calculando o rácio da estimativa do produto de coeficientes ab pelo respectivo erro padrão (Preacher e Hayes, 2004; MacKinnon, 2008; Preacher e Hayes, 2008;[2]). A conclusão final é obtida comparando o rácio com o valor crítico da distribuição Normal padrão, para um certo nível de significância α: se o valor absoluto da estatística de teste (o rácio) for superior ao quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição Normal reduzida, rejeita-se a hipótese nula de que o efeito indirecto é nulo (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008;[2]). Alternativamente é possível comparar o *p-value* do teste com o nível de significância: se o *p-value* for menor que α, o efeito indirecto é estatisticamente significativo (MacKinnon, 2008, Preacher e Hayes, 2008). Para a construção deste intervalo de confiança pressupõe-se que a dimensão da amostra utilizada para obtenção de a e b é suficientemente elevada.

Ambas as vias necessitam do cálculo do erro padrão de $\hat{a}\hat{b}$. Estimando as equações (1.2) e (1.3) obtêm-se as estimativas dos coeficientes a e b e os respectivos erros padrão, isto é, $s_{\hat{a}}$ e $s_{\hat{b}}$.[2] Os erros padrão são necessários para calcular o erro padrão do efeito indirecto:

$$s_{\hat{a}\hat{b}} = \sqrt{\hat{b}^2 s_{\hat{a}}{}^2 + \hat{a}^2 s_{\hat{b}}{}^2 + s_{\hat{a}}{}^2 s_{\hat{b}}{}^2} \quad (2.3)$$

(Preacher e Hayes, 2004; MacKinnon, 2008).

A expressão (2.4) corresponde à definição exacta do erro padrão do produto $\hat{a}\hat{b}$, no entanto, a versão frequentemente utilizada do teste de Sobel origina um valor aproximado do erro padrão do produto, omitindo o último termo da expressão (2.3):

$$s_{\hat{a}\hat{b}} = \sqrt{\hat{b}^2 s_{\hat{a}}{}^2 + \hat{a}^2 s_{\hat{b}}{}^2} \quad (2.4)$$

(Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; [2]).

2.2.2. Metodologia da diferença de coeficientes e respectivo teste de significância estatística

No contexto do modelo de mediação simples apresentado na secção 1.3, definir o efeito mediado como o produto ab ou $c - c'$ (diferença entre o efeito total e o efeito directo) é equivalente. Nestas circunstâncias, e para os mínimos quadrados ordinários e estimação de máxima verosimilhança, o efeito total é escrito como o somatório do efeito directo e do efeito indirecto. Caso a variável mediadora ou a variável resposta sejam dicotómicas, isto é, na regressão logística ou *probit*, a equivalência entre o produto ab e a diferença $c - c'$ não se verifica (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008;[2]). A metodologia da diferença de coeficientes, no contexto do modelo de mediação simples (secção 1.3), baseia-se na definição do efeito mediado como a subtracção dos coeficientes de regressão c e c' (MacKinnon, 2008; Newsom, 2015).

De notar, que no modelo de mediação múltipla o efeito indirecto é dado igualmente por $c - c'$, no entanto, as equações de regressão estimadas são diferentes, visto consistir num modelo com vários mediadores (MacKinnon, 2008; Preacher e Hayes, 2008).

A significância estatística do efeito mediado é possível de ser testada através das duas vias definidas na metodologia do produto de coeficientes (secção 2.2.1). O intervalo de confiança para o efeito indirecto (primeira via) é dado por:

$$\left((\hat{c} - \hat{c}') - z_{1-\frac{\alpha}{2}} * s_{\hat{c}-\hat{c}'} \, , (\hat{c} - \hat{c}') + z_{1-\frac{\alpha}{2}} * s_{\hat{c}-\hat{c}'} \right) \quad (2.5),$$

onde α representa o nível de significância, $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição Normal reduzida e $s_{\hat{c}-\hat{c}'}$ o erro padrão de $\hat{c} - \hat{c}'$ (MacKinnon, 2008).

A segunda via permite avaliar a significância estatística do efeito indirecto através da comparação de uma estatística de teste (um rácio) e o valor crítico de uma distribuição Normal padrão, para um dado nível de significância. O rácio apresenta como numerador a diferença entre os efeitos total e directo e como denominador o erro padrão da mesma diferença (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008). É esta a metodologia da diferença de coeficientes, proposta por Judd e Kenny (1981) (Newsom, 2015). A forma de conclusão é idêntica ao teste de Sobel: o efeito indirecto é estatisticamente significativo se, comparando o valor absoluto do rácio obtido com o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição Normal reduzida, para um certo nível de significância α, o primeiro for superior (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008) (Ou usando alternativamente a interpretação do *p-value*).

Qualquer uma das vias depende da obtenção do erro padrão de $c - c'$. O erro padrão da diferença dos coeficientes c e c' é dado por:

$$s_{\hat{c}-\hat{c}'} = \sqrt{s_{\hat{c}}{}^2 + s_{\hat{c}'}{}^2 - 2r s_{\hat{c}} s_{\hat{c}'}} \quad (2.6),$$

onde $rs_{\hat{c}}s_{\hat{c}'}$ representa a covariância entre c e c', e onde $s_{\hat{c}}$ e $s_{\hat{c}'}$ representam os erros padrões dos coeficientes de interesse estimados.

Visto as equações (2.4) e (2.6) serem semelhantes, a primeira é preferível por ser mais fácil de calcular (MacKinnon, 2008).

2.2.3. Metodologia de *Monte Carlo*

O método de *Monte Carlo* aplicado ao efeito indirecto foi proposto por MacKinnon, Lockwood e Williams (2004). Este método permite a construção de intervalos de confiança para o efeito indirecto através da sua distribuição empírica.[5] A distribuição de amostragem do efeito indirecto (como produto ab) considera as estimativas dos coeficientes de regressão a e b e os respectivos erros padrão, $s_{\hat{a}}$ e $s_{\hat{b}}$, resultantes da estimação das equações de regressão para a amostra original observada. Ou seja, através desses quatro valores, assumindo o modelo Normal, são geradas M amostras aleatórias resultantes da amostra original (e com a mesma dimensão desta), cada uma com parâmetros iguais a esses quatro valores iniciais. Para cada uma das M amostras são obtidas as respectivas estimativas de a e b, sendo possível obter ab e a sua distribuição de amostragem.

2.2.4. Metodologia *Bootstrap*

O método *Bootstrap* pertence à classe da metodologia de *Monte Carlo*, baseando-se na reamostragem com ou sem reposição, podendo ser um procedimento não paramétrico (Mackinnon,2008; Preacher e Hayes 2008; Yay, 2017;[2]). O método baseia-se na obtenção de amostras *bootstrap* com o objectivo de estimar um parâmetro de interesse. Ou seja, da amostra observada original de dimensão n começa-se por retirar uma amostra aleatória, com reposição, com a mesma dimensão, obtendo-se uma amostra *bootstrap* (MacKinnon, 2008). Este processo é repetido J vezes, obtendo-se J amostras *bootstrap*, onde J é um valor bastante elevado, implicando que seja um método computacionalmente intensivo (Preacher e Hayes, 2008; Yay, 2017;[4]).

O parâmetro de interesse, neste contexto, é o efeito indirecto, sendo a sua estimação efectuada pelo método *Bootstrap* não paramétrico. Consequentemente, não é necessário considerar a distribuição amostral do efeito indirecto como sendo uma distribuição Normal (metodologias do produto e da diferença de coeficientes) obtendo-se, ao invés, uma aproximação da distribuição, gerada empiricamente. A distribuição empírica é obtida reamostrando a amostra original J vezes, estimando-se em cada uma dessas reamostras, o efeito indirecto (Preacher e Hayes, 2008; Yay, 2017). Através da distribuição empírica para o efeito indirecto é possível obter o erro padrão deste efeito, o seu *p-value* e construir-lhe intervalos de confiança, os quais são assimétricos.[5] A estimativa de *bootstrap* do efeito indirecto corresponde à média da distribuição gerada, ou seja, à média das J estimativas obtidas do efeito indirecto (MacKinnon, 2008;[5]). O intervalo de confiança a $100 \times (1-\alpha)\%$ permite concluir sobre a presença ou não de mediação, sendo a forma mais utilizada o recurso aos percentis associados ao *bootstrap*, através dos quais se obtêm os percentis empíricos, correspondentes a $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$, como limites do IC (Yay, 2017).

2.2.5. Método da distribuição do produto

Um método alternativo de avaliação da significância estatística do efeito indirecto baseia-se na sua definição como o produto entre a e b. No contexto do modelo de mediação simples definido na secção 1.3, o produto ab é considerado como o produto de duas variáveis aleatórias normalmente distribuídas. No entanto, o produto de duas variáveis Normais não segue necessariamente uma distribuição Normal (MacKinnon, Lockwood e Williams, 2004; MacKinnon, Fairchild e Fritz, 2007;

MacKinnon, 2008). Consequentemente, Meeker, Cornwell e Aroian (1981) forneceram uma tabela para a distribuição do produto de duas variáveis aleatórias que seguem uma distribuição Normal padrão, visto o produto de variáveis com estas características não seguir uma distribuição que seja habitualmente usada. Com esta tabela é possível encontrar os valores críticos através dos valores populacionais $\delta_a = \frac{a}{\sigma_a}$ e $\delta_b = \frac{b}{\sigma_b}$, ou através dos equivalentes valores amostrais - que consistem numa aproximação aos valores populacionais -, $t_{\hat{a}} = \frac{\hat{a}}{s_{\hat{a}}}$ e $t_{\hat{b}} = \frac{\hat{b}}{s_{\hat{b}}}$. Alternativamente, é possível o cálculo directo dos valores críticos através do programa PRODCLIN, sendo necessária a introdução de dados, nomeadamente $\hat{a}, \hat{b}$, a sua correlação, $s_{\hat{a}}$, $s_{\hat{b}}$ e o nível de confiança desejado para o IC.

Obtidos os valores críticos, é possível construir o IC para o efeito indirecto, o qual é assimétrico, de forma a concluir sobre a presença ou ausência de mediação:

$$\left(\hat{a}\hat{b} + Valor\ crítico * s_{\hat{a}\hat{b}} ,\ \hat{a}\hat{b} + Valor\ crítico * s_{\hat{a}\hat{b}} \right) \quad (2.7)$$

(MacKinnon, Lockwood e Williams, 2004; MacKinnon, 2008).

2.2.6. Teste de significância conjunta

Uma forma alternativa de avaliar a significância estatística do efeito indirecto consiste em testar os coeficientes a e b separadamente como nulos, pois equivale a testar um efeito indirecto (na forma ab) nulo, visto tratar-se de um produto. Ou seja, o teste de significância conjunta não considera o primeiro passo da abordagem dos quatro passos, que afirma a necessidade de existir uma relação entre X e Y estatisticamente significativa.[5]

3. Modelo de mediação simples com variáveis binárias

Às metodologias apresentadas nos capítulos anteriores estão subjacentes o conjunto de equações (1.1), (1.2) e (1.3), definidas no contexto do modelo de mediação simples (secção 1.3). Neste capítulo são considerados, brevemente, os casos em que a variável resposta ou mediadora são binárias.

3.1. Variável resposta binária

Caso a variável resposta seja dicotómica, isto é, assume o valor 0 ou 1 (ausência ou presença de certa característica), não é possível estimar os efeitos através do modelo de mediação simples definido na secção 1.3, pois os pressupostos referentes às regressões lineares não são válidos, dificultando a estimação (MacKinnon, 2008).

O modelo para uma variável resposta dicotómica é estimado recorrendo à regressão *probit* ou logística. Atentando ao modelo de mediação simples habitual (secção 1.3), são as equações (1.1) e (1.2) que são estimadas por estas regressões, enquanto a equação (1.3) continua a ser estimada pelo método dos mínimos quadrados (MacKinnon, 2008).

Considerando a regressão logística, o modelo de mediação simples é dado por:

$$Y^* = \ln\left(\frac{p}{1-p}\right) = i_1 + cX + \varepsilon_1 \quad (3.1)$$

$$Y^* = \ln\left(\frac{p}{1-p}\right) = i_2 + c'X + bM + \varepsilon_2 \quad (3.2)$$

$$\mathrm{M} = i_3 + aX + \varepsilon_3 \quad (3.3),$$

onde p representa a probabilidade de sucesso, isto é, $p = P(Y = 1)$ e $\frac{p}{1-p}$ uma razão de probabilidades, nomeadamente a probabilidade de sucesso relativamente à de insucesso (*odds ratio*). Consequentemente, os coeficientes das regressões (3.1) e (3.2) vêm expressos no logaritmo dos *odds ratio*.

A regressão *probit* é idêntica à regressão logística, usando, ao invés da distribuição logística, a distribuição Normal. Com a regressão *probit*, a variável Y em (3.1) e (3.2) é reformulada como $Y^* = \Phi^{-1}(p)$, onde $\Phi(p)$ representa a função de distribuição da Normal padrão (MacKinnon, 2008).

É possível estimar o efeito indirecto através do produto e da diferença de coeficientes com uma variável resposta binária. A diferença relativamente às metodologias apresentadas nas secções 2.2.1 e 2.2.2 consiste no facto do produto e da diferença de coeficientes não fornecerem valores iguais para o efeito indirecto, devido à variância dos erros. A variância dos erros é mantida fixa nas regressões *probit* e logística, respectivamente, em 1 e $\frac{\pi^2}{3}$, pois a variável dependente não é directamente observável, ao invés do que acontece nos modelos de regressão linear, na qual a variância da variável dependente é observável e constante (MacKinnon,1993; MacKinnon, 2008). Consequentemente, é necessária a estandardização das estimativas das equações (3.1) e (3.2) antes de ser avaliada a mediação para que a igualdade se verifique, isto é, $ab = c - c'$. A estandardização corresponde a dividir a estimativa de c pelo desvio padrão (σ) do Y^* na equação (3.1), e as estimativas de c' e b pelo desvio padrão (σ) do Y^* na equação (3.2). Estas quantidades são dadas por:

$$\hat{\sigma}^2_{Y^*} = \hat{c}^2\hat{\sigma}^2_X + W \quad (3.4)$$

$$\hat{\sigma}^2_{Y^*} = \widehat{c'}^2\hat{\sigma}^2_X + \hat{b}^2\hat{\sigma}^2_M + 2\widehat{c'}\hat{b}\hat{\sigma}_{XM} + W \quad (3.5),$$

onde (3.4) é relativa à equação (3.1) e (3.5) é relativa à equação (3.2). Na regressão logística $W = \frac{\pi^2}{3}$ e na regressão *probit* $W = 1$ (MacKinnon,1993; MacKinnon, 2008).

3.2 Variável mediadora binária

Caso a variável mediadora seja binária provoca uma reformulação na equação (1.3), a qual vai ser estimada pela regressão logística ou *probit*, enquanto as restantes equações continuam a ser estimadas pelo *OLS*. Considerando a regressão logística, o modelo de mediação simples é dado por:

$$\mathrm{Y} = i_1 + cX + \varepsilon_1 \quad (3.6)$$

$$\mathrm{Y} = i_2 + c'X + bM + \varepsilon_2 \quad (3.7)$$

$$M^* = logit\left(\frac{p}{1-p}\right) = i_3 + aX + \varepsilon_3 \quad (3.8),$$

considerando o modelo *probit* $M^* = \Phi^{-1}(p)$ (MacKinnon, 2008).

Com um mediador binário, para que o produto e a diferença de coeficientes gerem uma estimativa semelhante para o efeito indirecto efectua-se a estandardização da estimativa a. Isto é, obtém-se o quociente entre a estimativa a e $\hat{\sigma}_{M^*}$ da equação (3.8), o qual é dado por:

$$\hat{\sigma}^2_{M^*} = \hat{a}^2\hat{\sigma}^2_X + W \quad (3.9),$$

com $W = \frac{\pi^2}{3}$ na regressão logística e $W = 1$ na regressão *probit*.

4. Abordagem contrafactual à análise de mediação causal

A abordagem contrafactual baseia-se no conceito de resultados potenciais, consistindo numa alternativa à abordagem padrão (Imai, Tingley e Keele, 2010; Imai, Tingley e Yamamoto, 2010; Qin, 2016). A razão está no facto da abordagem padrão conter limitações, nomeadamente a inexistência de uma definição geral para os efeitos de mediação causal, variando com os modelos estatísticos usados; a incapacidade de especificar a hipótese fundamental da identificação; e a dificuldade de generalização a modelos não lineares, tais como modelos logísticos e *probit*, a modelos semi-paramétricos e não paramétricos e a variáveis mediadoras e de resultado de suporte discreto (Imai, Tingley e Keele, 2010).

4.1. Introdução ao Modelo Causal de Resultados Potenciais

Nesta abordagem, o conceito de resultados potenciais é fundamental na definição de causalidade e, consequentemente, na definição dos efeitos causais. Primeiramente é necessário apresentar alguns conceitos introdutórios essenciais (Abadie, 2005).

Subjacente à causalidade está um par acção – unidade, isto é, a causalidade está relacionada com uma acção que vai ser aplicada a uma unidade (indivíduo, objecto físico, conjunto de objectos ou indivíduos, etc.), num dado momento. Com isto, é possível introduzir a primeira variável fundamental neste contexto: o tratamento. O tratamento é uma variável que assume dois ou mais valores, conforme a existência de duas ou mais acções. Normalmente é uma variável binária, sendo a acção activa designada de tratamento activo ou tratamento e a acção passiva designada de tratamento de controlo ou controlo (Imbens e Rubin, 2015).

A variável tratamento pode ser definida, para a unidade amostral i, como:

$$T_i = \begin{cases} 1, & se\ a\ unidade\ i\ recebe\ o\ tratamento \\ 0, & caso\ contrário \end{cases} \quad (4.1).$$

A segunda variável fundamental para a definição de resultados potenciais é a variável resposta ou resultado, a qual é denotada, para a unidade amostral i, como Y_i (Abadie, 2005). Por fim, pode-se introduzir a notação X_i, representando, para cada unidade amostral i, o conjunto de covariáveis observáveis que, para além da variável tratamento, influenciam a variável resultado (Esarey, 2015). Como estas covariáveis apresentam uma precedência temporal relativamente à variável tratamento, alguns autores designam-nas como covariáveis de pré-tratamento observáveis (VanderWeele e Vansteelandt, 2009; Imai, Keele e Yamamoto, 2010). De acordo com a secção 1.5.3, as covariáveis são consideradas factores de confundimento a partir do momento em que confundem a relação causal X-Y, daí que outros autores se refiram a X_i como factores de confundimento de pré-tratamento observáveis (Abadie, 2005; Imai, Tingley e Keele, 2010).

Fazendo a analogia com o modelo de mediação simples definido anteriormente, a variável tratamento corresponde à variável independente X enquanto a variável resposta corresponde à variável dependente Y.

4.1.1. Pressuposto *SUTVA*

Um pressuposto a considerar para obter uma correcta definição dos resultados potenciais, e consequentemente dos efeitos causais, é o pressuposto *SUTVA*, o qual deriva do inglês *Stable Unit Treatment Value Assumption* (Morgan e Li, 2014). A definição correcta dos resultados potenciais relaciona-se com as duas propriedades constituintes do *SUTVA*: a propriedade de não interferência (entre unidades amostrais) e a propriedade de variações inesperadas dos níveis de tratamentos (Qin,

2016; Imbens e Rubin, 2015). A primeira refere que os "resultados potenciais não variam, para qualquer unidade amostral, com os tratamentos atribuídos a outras unidades amostrais" (Qin, 2016: p.15); a segunda refere que "para cada unidade amostral, não existem formas ou versões diferentes de cada nível de tratamento" (Qin, 2016: p.15), ou seja, "para cada unidade amostral, existe apenas uma única versão de cada nível de tratamento" (Morgan e Li, 2014: p.1).

A grande limitação do *SUTVA* consiste no facto de ao se tratar de um pressuposto, implica que seja efectuado um estudo cuidadoso, caso contrário, apesar de possível, a inferência causal não é totalmente fiável (Morgan e Li, 2014). Em certas situações, pode-se verificar uma interferência entre as várias unidades amostrais, provocando que o resultado potencial para uma unidade amostral dependa do tratamento atribuído a outra unidade amostral. Por outro lado, o resultado potencial de uma unidade amostral pode sofrer variações ao receber o mesmo nível de tratamento com outras versões desse tratamento (Imbens e Rubin, 2015). Ou seja, em certas situações, pode ocorrer uma violação das duas propriedades constituintes do *SUTVA*, logo assume-se este pressuposto no caso de existir um conhecimento prévio do caso em estudo (Imbens e Rubin, 2015; Morgan e Li, 2014).

Imbens e Rubin (2015) exemplificam a importância do *SUTVA*, através da propriedade de não interferência, na correcta definição dos resultados potenciais. Para tal, é considerada uma variável tratamento com acção activa correspondendo a "tomar uma aspirina" e acção passiva correspondendo a "não tomar uma aspirina"; e uma variável resposta binária - "melhoria de dor de cabeça" e "não melhoria da dor de cabeça". Não assumindo o *SUTVA*, existiriam 4 níveis de tratamento no total: "o indivíduo 1 toma aspirina e o indivíduo 2 não", "o indivíduo 1 não toma aspirina e o indivíduo 2 toma", "nenhum indivíduo toma aspirina" e "os dois indivíduos tomam aspirina"; e consequentemente, quatro resultados potenciais para cada indivíduo, um para cada nível de tratamento. No entanto, baseando-se em conhecimento prévio, constata-se, por exemplo, que o indivíduo 1 ao tomar uma aspirina, não implica uma melhoria da dor de cabeça do indivíduo 2. Ou seja, neste exemplo, os resultados potenciais de uma unidade amostral não são função do tratamento atribuído a outra unidade.

4.1.2. Definição de resultados potenciais

Admitindo o pressuposto *SUTVA*, e considerando a variável tratamento como dicotómica, é possível definir resultados potenciais ou contrafactuais (VanderWeele e Vansteelandt, 2009), os quais se referem a possíveis "casos contrafactuais", isto é, a situações que não aconteceram mas poderiam ter acontecido (Abadie, 2015: p.260). No exemplo de Imbens e Rubin (2015), apresentado na secção 4.1.1, o resultado potencial para um indivíduo que tome uma aspirina pode ser a melhoria da dor de cabeça ou a não melhoria.

O resultado potencial é definido para cada nível de tratamento. Neste sentido, uma variável tratamento dicotómica provoca a existência de dois resultados potenciais, para cada unidade amostral (VanderWeele, 2016): Y_i (1) representa o resultado potencial sob $T_i = 1$ e $Y_i(0)$ representa o resultado potencial sob $T_i = 0$ (Abadie, 2005; Qin, 2016). Pelo exemplo de Imbens e Rubin (2015), apresentado na secção 4.1.1, cada indivíduo tem como possíveis acções "tomar uma aspirina" ou "não tomar uma aspirina" (variável tratamento dicotómica) e como resposta, uma variável que assume dois valores - "melhoria da dor de cabeça" e "não melhoria da dor de cabeça" (variável resposta dicotómica). Como uma variável tratamento binária provoca a existência de dois resultados potenciais para cada unidade amostral, Y_i("tomar uma aspirina") representa o estado da dor de cabeça para o indivíduo i se este tomar uma aspirina, isto é, verifica-se uma melhoria ou não; Y_i("não tomar uma aspirina") representa o estado da dor de cabeça para o indivíduo i se este não tomar uma aspirina (melhoria ou não da dor de cabeça).

A priori os dois resultados potenciais são possíveis de serem observados, conforme a unidade amostral receba o nível de tratamento correspondente (Imbens e Rubin, 2015). Contudo, somente um desses resultados potenciais é realizado e observado, aquele correspondente ao nível de tratamento observado, isto é, à acção tomada. O outro resultado potencial (ou outros, caso a variável tratamento assuma mais que dois valores) é impossível de observar, visto a correspondente acção não ter sido tomada (Abadie, 2015; Imbens e Rubin, 2015). O resultado realizado e observado, para a unidade amostral i, é dado por:

$$\mathrm{Y}_i^{obs} = Y_i(T_i) = \begin{cases} Y_i(1), & se\ T_i = 1 \\ Y_i(0), & se\ T_i = 0 \end{cases} \quad (4.2).$$

Admitindo o *SUTVA*, os valores possíveis para a variável resultado, e consequentemente o resultado potencial realizado e observado depende apenas da acção individual de cada unidade amostral (Imbens e Rubin, 2015; Imai, Tingley e Keele, 2010). A expressão (4.2) é equivalente a:

$$\mathrm{Y}_i^{obs} = T_i \,.\, Y_i(1) + (1 - T_i).\, Y_i(0) \quad (4.3)$$

(Abadie, 2005).

O resultado potencial não realizado e não observado é dado por:

$$Y_{\ i}^{\ mis} \ = \ Y_i(1 - T_i) = \begin{cases} Y_i(1), & se\ T_i = 0 \\ Y_i(0), & se\ T_i = 1 \end{cases} \quad (4.4)$$

(Imbens e Rubin, 2015).

Retomando o exemplo anterior, caso o nível de tratamento observado para um indivíduo ser "tomar uma aspirina", isto é, caso seja esta a acção tomada ($T_i = 1$), o resultado potencial correspondente - Y_i("tomar uma aspirina") $= Y_i(1)$ – é observado, ou seja, ao tomar uma aspirina verifica-se uma melhoria na dor de cabeça ou não. No entanto, é impossível saber o que aconteceria se o indivíduo optasse por não tomar a aspirina, isto é, se haveria uma melhoria ou não da dor de cabeça, pois este optou por tomar o medicamento ($Y_i(0)$ não é observado sob $T_i = 1$).

4.1.3 Mecanismo de atribuição

O mecanismo de atribuição consiste num tópico fundamental na definição e identificação dos efeitos causais, remetendo para a forma como a atribuição das unidades amostrais a cada tipo de tratamento é realizada. Consequentemente estabelece os resultados potenciais observados e não observados para cada unidade amostral (Rubin, 2005; Imbens e Rubin, 2015). Nos estudos experimentais a atribuição do tratamento a cada unidade amostral é aleatória, não o sendo nos estudos não experimentais (Abadie, 2005; Rubin, 2005; Esarey, 2015; Imbens e Rubin, 2015).

4.1.4. Identificação e estimação dos efeitos causais

4.1.4.1. Efeito causal do tratamento

O efeito causal do tratamento sobre a resposta é definido como a diferença entre o resultado potencial quando o tratamento é aplicado comparativamente ao mesmo resultado na ausência da sua aplicação, ou seja, consiste na diferença entre os dois resultados potenciais, para cada unidade amostral i, em igual momento:

$$Y_i(1) - Y_i(0) \quad (4.5)$$

(Abadie, 2005; Esarey, 2015; Qin, 2016).

Denotando por X um vector de covariáveis relacionados com a resposta, para além da variável tratamento, o efeito causal do tratamento sobre a resposta é dado por:

$$Y_i(T = 1, X) - Y_i(T = 0, X) \quad (4.6)$$

(Esarey, 2015).

O efeito causal do tratamento resulta da comparação dos resultados potenciais, no mesmo momento, para a mesma unidade amostral. Por conseguinte, uma mesma unidade amostral em momentos diferentes é considerada uma outra unidade, pois a ideia subjacente é que o efeito causal do tratamento sobre a resposta se deva unicamente ao tratamento. Isto é, o objectivo é realizar uma comparação entre unidades idênticas, em termos de características (covariáveis X), considerando que embora a atribuição de uma unidade seja a um dado nível de tratamento, poderia ter sido ao nível de tratamento oposto. Entre observações diferentes, a unidade amostral sofre alterações nas suas características, provocando uma comparação de unidades heterogéneas (Esarey, 2015; Imbens e Rubin, 2015).

A definição do efeito é independente do resultado potencial observado (Imbens e Rubin, 2015), contrariamente à sua estimação e inferência. Não é possível estimar o efeito causal do tratamento para a mesma unidade amostral, devido à impossibilidade de observar simultaneamente a mesma unidade amostral sob cada condição de tratamento (tratamento ou controlo) (Abadie, 2005; Imai, Tingley e Keele, 2010; Esarey, 2015) - "problema fundamental da inferência causal" segundo Holland (1986: p.947) (Imbens e Rubin, 2015). A resolução do problema consiste na consideração de várias unidades amostrais, umas que recebem o tratamento activo e outras que recebem o tratamento de controlo, conforme o mecanismo de atribuição e o pressuposto *SUTVA* (Imbens e Rubin, 2015).

O efeito causal como comparação dos resultados potenciais surge frequentemente em situações da vida quotidiana. O filme "It's a Wonderful Life" aborda estes conceitos. Neste filme, a personagem principal, interpretado por James Stewart como George Bailey, enfrenta uma depressão, acreditando que o mundo seria melhor se este não tivesse nascido, ponderando o suicídio. Num momento oportuno um anjo aparece, mostrando-lhe as consequências que ocorreriam se ele não tivesse nascido, nomeadamente, não teria os seus filhos, o seu irmão morreria afogado porque George não o teria conseguido salvar, entre outros inconvenientes. Perceber a importância que tem na vida das pessoas em seu redor faz com que George desista da ideia de suicídio. Fazendo a analogia com o tema de estudo, verifica-se que o mundo real é o resultado realizado e observado sob o estado de tratamento "George nasceu", ou seja, representa os acontecimentos observados na vida de George por este ter nascido; o mundo mostrado pelo anjo (mundo contrafactual) representa o resultado potencial não observado, sob o estado de tratamento "George nasceu", isto é, representa os acontecimentos na vida de George por este não ter nascido, os quais não são possíveis de observar, pois na realidade George nasceu. O efeito causal de George não nascer consiste na comparação dos dois resultados potenciais, isto é de todos os acontecimentos que aconteceram na vida de George, com todos aqueles que aconteceriam caso George não tivesse nascido (Imbens e Rubin, 2015).

4.1.4.2. Pressuposto de ignorabilidade forte

Além do pressuposto *SUTVA*, outro pressuposto necessário à identificação, ou seja, à estimação consistente dos efeitos causais, é o pressuposto de ignorabilidade da atribuição do tratamento. A sua definição depende do tipo de dados em estudo, se são dados provenientes de estudos experimentais ou dados de estudos não experimentais ou observáveis.

Num estudo experimental, a atribuição do tratamento a cada unidade amostral, retirada aleatoriamente de uma dada população, é aleatória. Consequentemente, a atribuição de uma condição da variável tratamento a uma unidade amostral i é independente dos resultados possíveis da mesma unidade

amostral, isto é, o tratamento é ignorável (Imai, Tingley e Keele, 2010; Esarey, 2015). Neste tipo de estudo é possível controlar qualquer covariável que possa influenciar a relação tratamento-resposta (factor de confundimento), permitindo concluir que a diferença na variável resposta no grupo de tratamento e controlo se deve unicamente ao tratamento aplicado (Esarey, 2015). Neste caso, o pressuposto de ignorabilidade é mantido:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \quad (4.7)$$

(Abadie, 2005; Imai, Tingley e Keele, 2010; Esarey, 2015; Qin, 2016).

Num estudo não experimental é necessário um pressuposto mais forte (Esarey, 2015; Qin, 2016). Ao contrário dos estudos experimentais, normalmente não se atribui aleatoriamente uma unidade amostral a cada grupo (de tratamento ou de controlo) tornando a expressão (4.7) insuficiente (Abadie, 2005; Esarey, 2015). Na presença de dados observados é provável a existência de diferentes características entre as unidades amostrais, susceptíveis de afectar a relação tratamento – resposta (factor de confundimento), pelo que deve ser controlado o maior número de características observáveis (Abadie, 2005). O pressuposto de ignorabilidade forte é dado por:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i | X_i = x \quad (4.8)$$

(Abadie, 2005; Esarey, 2015; Qin, 2016).

A expressão (4.8) supõe o controlo das covariáveis de pré-tratamento possíveis de influenciar a relação entre a variável tratamento e a variável resposta (factores de confundimento), provocando uma independência estatística entre tratamento e resultado. Com isto, é possível comparar os valores da variável resposta entre unidades amostrais, sujeitas a cada tipo de tratamento, permitindo uma comparação entre unidades com as mesmas características, isto é, uma comparação entre grupos homogéneos. Ou seja, na prática, este pressuposto permite afirmar que o efeito causal do tratamento na resposta se deve unicamente ao tratamento e não a outras covariáveis.

4.1.4.3. Efeito do tratamento médio

O interesse na definição do efeito do tratamento médio (*ATE*, do inglês *Average Treatment Effect)* é motivado pelo problema de inferência causal (Abadie, 2005; Imai, Tingley e Keele, 2010; Esarey, 2015). A sua definição depende do mecanismo de atribuição do tratamento às unidades amostrais (secção 4.1.3), ou seja, varia consoante se esteja perante um estudo experimental ou não experimental. É possível identificar o efeito do tratamento médio considerando os pressupostos *SUTVA* e ignorabilidade do tratamento.

Num estudo experimental, o efeito do tratamento médio, para cada unidade i, é dado por:

$$E[Y_i(1) - Y_i(0)] \quad (4.9).$$

Ou seja, consiste na aplicação do valor esperado sobre o efeito causal do tratamento, através de uma amostra extraída aleatoriamente de uma certa população (Imai, Tingley e Keele, 2010).

Para a população em estudo, o efeito é dado por:

$$E[Y(1) - \mathrm{Y}(0)] \quad (4.10)$$

(Abadie 2005; VanderWeele e Vansteelandt, 2009; Qin, 2016).

Assumindo o pressuposto de independência estatística entre o tratamento e os resultados potenciais (ignorabilidade do tratamento), o efeito do tratamento médio para a população é identificado por:

$$E[Y(1) - Y(0)] = E[Y|T = 1] - E[Y|T = 0] \quad (4.11)$$

(Abadie, 2005; Qin, 2016).

Num estudo não experimental, considerando as possíveis covariáveis observadas para as unidades amostrais, o efeito do tratamento médio, para cada unidade i, é dado por:

$$E\,[Y_i(T = 1, X) - Y_i(T\ =\ 0, X)] =\ E\,[Y_i(T\ =\ 1, X)] -\ E\,[Y_i(T\ =\ 0, X)] \quad (4.12)$$

(Esarey, 2015).

Para a população em estudo, o efeito é dado por:

$$E\,[Y(T = 1, X) - Y(T\ =\ 0, X)] =\ E\,[Y(T\ =\ 1, X)] -\ E\,[Y(T\ =\ 0, X)] \quad (4.13)$$

Sob o pressuposto de ignorabilidade forte do tratamento, o efeito do tratamento médio para a população é identificado por:

$$E[Y(1) - Y(0)|X] = E[Y|T = 1, X] - E[Y|\,T = 0, X] \quad (4.14)$$

(Abadie, 2005; Qin, 2016).

Como o objectivo é o estudo da análise de mediação causal no contexto dos resultados potenciais, não irão ser aprofundadas as técnicas de estimação dos efeitos causais.

4.2. Extensão dos resultados potenciais aos efeitos de mediação causal

Com a introdução à abordagem contrafactual é possível estender o conceito de resultados potenciais à análise de mediação causal. A mediação, como referido, remete para a existência de uma ou mais variáveis mediadoras entre a variável tratamento e a variável resposta. Especificamente, na presença de mediação, T causa M o qual causa Y. Por conseguinte, são definidos mediadores potenciais para cada estado de tratamento. Como a variável tratamento é na maioria dos casos dicotómica, os mediadores potenciais correspondem a $M_i(1)$ e $M_i(0)$, denotando i a unidade amostral. No entanto, apenas um é possível de ser observado: aquele que tem o correspondente tratamento observado. O mediador potencial observado é dado por $M_i = M_i(T_i)$, isto é, sob o estado de tratamento $T_i = t$, com $t = 0{,}1$, é possível observar o valor do mediador sob t mas não o valor do mediador sob $T_i = 1 - t$ (Imai, Tingley e Keele, 2010).

Por exemplo, considere-se uma variável tratamento binária $(T = 0, 1)$ e uma variável mediadora binária, a qual assume o valor A ou o valor B, para um certo indivíduo. Caso o indivíduo receba o tratamento activo $(T = 1)$ apenas $M(1)$ é observável, isto é, é possível observar se o valor da mediadora é A ou B, sob esse tratamento. No entanto, caso o indivíduo receba o tratamento activo, não é possível observar $M(0)$, ou seja, o valor da mediadora (A ou B) caso o indivíduo tivesse recebido o tratamento de controlo.

Na secção 4.1.2, definiram-se os resultados potenciais como função do estado da variável tratamento. No contexto da mediação, os resultados potenciais dependem do estado de tratamento e da variável

mediadora, visto a variável resposta ser influenciada tanto pelo tratamento como pelo mediador (Imai, Tingley e Keele, 2010). Sob o estado de tratamento $T_i = t$, o resultado potencial observado para a unidade amostral i é dado por $Y_i = Y_i(\mathrm{t}, M_i(t))$, não sendo possível observar $Y_i(\mathrm{t}, M_i(1-t)$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010).

De seguida, são definidos, no contexto da mediação, os efeitos causais de nível unitário, os quais não podem ser estimados devido ao problema de inferência causal, recorrendo-se à estimação dos respectivos efeitos médios.

4.2.1. Efeito de mediação causal

O efeito de mediação causal ou efeito indirecto (natural, segundo Pearl (2001)) sob o estado de tratamento t, para cada unidade amostral i, define-se como a diferença entre os resultados potenciais sob o estado de tratamento t. Ou seja, representa, sob o estado de tratamento t, com t a assumir o valor 0 ou 1, a diferença na variável resultado se se alterasse o valor do mediador sob o tratamento de controlo para o valor do mediador sob o tratamento activo. Consequentemente, surge o interesse na definição da variável mediadora, pois o mediador potencial pode assumir valores diferentes nos dois estados da variável tratamento. De acordo com Pearl (2001), o efeito de mediação causal corresponde ao efeito de M no caminho causal de T sobre Y. O efeito de mediação causal, para cada unidade amostral i, é dado por:

$$\delta_i(t) \equiv Y_i\big(t, M_i(1)\big) - Y_i\big(t, M_i(0)\big) \qquad (4.15),$$

com $t = 0{,}1$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010).

Não é possível estimar o efeito de mediação causal pois, sob o estado de tratamento t, é possível observar $Y_i\big(t, M_i(t)\big)$ mas não $Y_i\big(t, M_i(1-t)\big)$. Ou seja, para o grupo de tratamento ($t = 1$), apenas é possível observar $Y_i\big(t, M_i(1)\big)$, enquanto para o grupo de controlo ($t = 0$) é possível somente observar $Y_i\big(t, M_i(0)\big)$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010). De acordo com Robins (2003), $\delta_i(1)$ representa o efeito indirecto total, enquanto $\delta_i(0)$ representa o efeito indirecto puro (Imai, Keele e Yamamoto, 2010).

O efeito de mediação causal é nulo caso o valor tomado pela variável tratamento não influencie o valor do mediador, isto é, $M_i(1) = M_i(0)$ (Imai, Tingley e Keele, 2010).

4.2.2. Efeito médio de mediação causal

Considerando a definição de efeito de mediação causal na secção 4.2.1., define-se o efeito médio de mediação causal (*ACME*, do inglês *Average Causal Mediation Effect*) como:

$$\bar{\delta}(t) = E\big(\delta_i(t)\big) \equiv E\big[Y_i\big(t, M_i(1)\big) - Y_i\big(t, M_i(0)\big)\big] \quad (4.16),$$

com $t = 0{,}1$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010, Qin, 2016).

4.2.3. Efeito directo

Segundo Pearl (2001), é necessário distinguir entre duas variantes do efeito directo: o efeito directo natural do tratamento e o efeito directo controlado do tratamento (Imai, Tingley e Keele, 2010).

4.2.3.1. Efeito directo natural ou puro

O efeito directo natural do tratamento sobre o resultado (vulgarmente designado como efeito directo) para cada unidade amostral i, é dado por:

$$\zeta_i\,(t) \equiv Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big) \quad (4.17),$$

com $t = 0{,}1$. Ou seja, representa o efeito do tratamento sobre o resultado, mantendo constante o mediador no nível t (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010).

De acordo com Robins (2003), $\zeta i\,(1)$ representa o efeito directo total e $\zeta i\,(0)$ representa o efeito direto puro (Imai, Keele e Yamamoto, 2010).

Identicamente ao efeito de mediação causal, apenas é possível observar, sob o estado de tratamento t, $Y_i\big(t, M_i(t)\big)$, razão pela qual é necessário considerar a estimação do correspondente efeito médio. Apesar deste aspecto comum, a expressão (4.15), referente ao efeito de mediação causal difere da expressão (4.17). Na primeira, o objectivo é verificar se, para um dado nível de tratamento, uma alteração no valor da variável mediadora de um estado de tratamento para o alternativo causa diferenças na variável resposta; na segunda expressão ocorre o oposto: o valor do mediador mantém-se fixo sob um dado nível de tratamento, sendo o objectivo verificar a diferença na variável resposta causada pela alteração do valor da variável tratamento.

4.2.3.2. Efeito directo controlado do tratamento

Segundo Pearl (2001) e Robins (2003) o efeito directo controlado do tratamento, para a unidade amostral i, é dado por**:**

$$Y_i(1, m) - Y_i(0, m) \quad (4.18).$$

Ou seja, a diferença relativamente à expressão (4.17) consta no valor assumido pela variável mediadora, o qual é fixo na expressão (4.18), não dependendo do grupo de tratamento da unidade amostral (Imai, Keele e Yamamoto, 2010). Além disso, implícito ao efeito directo controlado está o efeito directo controlado do mediador, o qual é possível comparar com o efeito de mediação causal definido na expressão (4.15). O efeito directo controlado do mediador, para a unidade amostral i, define-se como**:**

$$Y_i(t, m) - Y_i(t, m') \quad (4.19),$$

para $t = 0{,}1$, $m \neq m'$ (Imai, Keele e Yamamoto, 2010; Qin, 2016). Comparando as expressões (4.15) e (4.19), verifica-se que a diferença consta na variável mediadora: ao invés de $Mi(1)$ e $Mi(0)$ – valores potenciais desta variável - estão valores concretos. No entanto, os efeitos directos controlados (do tratamento e do mediador) são importantes no contexto de modelos de moderação e não no contexto de mediação (Imai, Keele e Yamamoto, 2010).

4.2.4. Efeito directo natural médio

Aplicando o valor esperado à expressão (4.17) obtém-se o efeito directo (natural) médio (*ADE*, do inglês *Average Direct Effect)*:

$$\bar{\zeta}(t) = E\big(\zeta_i(t)\big) \equiv E\big[Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big)\big] \quad (4.20),$$

para $t = 0{,}1$ (Imai, Tingley e Keele, 2010; Qin, 2016).

4.2.5. Efeito total

O efeito causal total da variável tratamento sobre a variável resposta, no contexto da mediação, para cada unidade amostral i, é definido como a soma do efeito de mediação causal sob o nível de tratamento t e o efeito directo (natural) sob $1-t$:

$$\tau_i(t) = \delta_i(t) + \zeta_i(1-t) \quad (4.21),$$

para $t = 0{,}1$ (Imai, Keele e Yamamoto, 2010).

Uma simplificação da expressão (4.21) é possível, pois independentemente do valor de $t, t = 0{,}1$, o efeito total é dado por:

$$\tau_i \equiv Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(0)\big) \quad (4.22)$$

(Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010).

Assumindo a independência do efeito directo e do efeito de mediação causal relativamente ao valor da variável tratamento (pressuposto de não interacção), para $t = 0{,}1$:

$$\delta_i(t) = \delta_i(1-t) = \delta_i \quad (4.23)$$

$$\zeta_i(t) = \zeta_i(1-t) = \zeta_i \quad (4.24).$$

Consequentemente, o efeito total pode ser escrito simplificadamente como:

$$\tau_i = \delta_i + \zeta_i \quad (4.25).$$

Ou seja, sob esta hipótese, o efeito total é composto pela soma do efeito mediação causal e do efeito directo (Imai, Tingley e Keele, 2010).

4.2.6. Efeito total médio

Considerando a expressão (4.21), o valor médio do efeito total é dado por:

$$\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1-t) \quad (4.26),$$

para $t = 0{,}1$ (Imai, Keele e Yamamoto, 2010).

Por simplificação, $\forall\, t, t = 0{,}1$:

$$\bar{\tau} = E(\tau_i) \equiv E\big[Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(0)\big)\big] \quad (4.27),$$

(Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010).

Analogamente a (4.25), é válido escrever:

$$\bar{\tau} = E(\delta_i + \zeta_i) = \bar{\delta} + \bar{\zeta} \quad (4.28),$$

sob o pressuposto mencionado de não interacção, o qual, para $t = 0{,}1$, refere:

$$\bar{\delta}(t) = \bar{\delta}(1-t) = \bar{\delta} \quad (4.29)$$

$$\bar{\zeta}(t) = \bar{\zeta}(1-t) = \bar{\zeta} \quad (4.30).$$

Ou seja, sob esta hipótese, o efeito total é composto pela soma do efeito mediação causal médio e do efeito directo médio (Imai, Tingley e Keele, 2010).

4.3. Identificação dos efeitos

4.3.1. Hipótese de ignorabilidade sequencial

Numa análise de mediação os efeitos que se destacam dos efeitos anteriormente descritos, sendo necessário estimá-los, são: o efeito de mediação causal médio (*ACME*), o efeito directo médio (ADE) e o efeito total médio. O *ACME* é o que tem mais interesse de entre os restantes. De forma a estimar os efeitos de forma consistente e não paramétrica, é necessário verificar uma suposição fundamental: a ignorabilidade sequencial.

A hipótese de ignorabilidade sequencial é definida como:

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \,|X_i = x, \quad (4.31)$$
$$Y_i(t', m) \perp\!\!\!\perp M_i(t)|T_i = t, X_i = x, \quad (4.32)\}$$

com $t, t' = 0{,}1$ e x pertencente ao suporte da distribuição de X_i, $\mathcal{X}$.
Para além disso as probabilidades condicionais $P(T_i = t\,|X_i = x)$ e $P(M_i(t) = m|T_i = t, X_i = x)$ pertencem ao intervalo entre 0 e 1, para $t = 0{,}1$, $x \in \mathcal{X}$ e m pertence ao suporte da distribuição de M_i, $\mathcal{M}$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010; Qin, 2016).

A hipótese de ignorabilidade sequencial é designada desta forma pois é composta por duas expressões de ignorabilidade apresentadas sequencialmente, nomeadamente a ignorabilidade do tratamento e a ignorabilidade do mediador. A ignorabilidade do tratamento – expressão (4.31) -, remete para o facto de, para cada unidade amostral i, o tratamento dado os fatores de confundimento/covariáveis de pré – tratamento observados, $X_i = x$, ser independente quer dos valores dos resultados potenciais quer dos valores possíveis que a variável mediadora pode tomar, $\{Y_i\,(t', m), M_i(t)\}$ (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010; Qin, 2016). A ignorabilidade do tratamento é condição suficiente para identificar o efeito total médio (Tingley et al., 2014). A ignorabilidade do mediador – expressão (4.32) - afirma que, para cada unidade amostral i, dado as covariáveis de pré-tratamento observadas, $X_i = x$, e a variável de tratamento observada, isto é, assumindo um dado valor t ($t = 0{,}1$), o mediador é independente dos valores possíveis da variável resultado, isto é, é ignorável (Imai, Keele e Yamamoto, 2010; Imai, Tingley e Keele, 2010; Qin, 2016). A ignorabilidade do mediador é mais forte relativamente à ignorabilidade do tratamento devido à possibilidade de existência dos factores de confundimento, mas não de pré-tratamento e observáveis. Pelo conceito introduzido de X_i, verifica-se que o condicionamento na equação (4.32), tal como em (4.31), é realizado nas covariáveis que precedem temporalmente o tratamento e não nas covariáveis que o sucedem. Por conseguinte, a relação mediador-resultado pode ficar susceptível ao aparecimento de covariáveis (de pós-tratamento) mensuráveis ou não mensuráveis ou a covariáveis de pré-tratamento não medidas (Imai, Tingley e Keele, 2010; Tingley et al., 2014).

Em estudos experimentais verifica-se a condição (4.31), pois a atribuição do tratamento a cada unidade amostral é aleatória. No entanto, a condição (4.32) pode não se verificar, visto ser mais forte comparativamente a (4.31) (Imai, Tingley e Keele, 2010).

Nos estudos não experimentais não se garante que a condição (4.31) seja satisfeita, pois neste tipo de estudos as unidades amostrais não são atribuídas ao acaso para o grupo de tratamento ou controlo, podendo sim auto-seleccionar-se no grupo. A conclusão é idêntica para (4.32).

Devido às limitações apresentadas e, particularmente pela dificuldade de verificação da hipótese de ignorabilidade sequencial em estudos não experimentais (observacionais), são aplicadas várias técnicas de análise de sensibilidade (Imai, Tingley e Keele, 2010).

4.3.2. Exemplo de análise de mediação causal

Um exemplo de mediação causal é o estudo experimental apresentado por Nelson, Clawson e Oxley (1997). Este estudo consistiu em verificar a influência da forma de abordagem dos meios de comunicação social, em questões relacionadas com política, sobre a formação de opinião de um indivíduo, admitindo que a existência de diferentes meios de comunicação social implica a existência de diferentes formas de abordagem da mesma notícia. O estudo teve como amostra um conjunto de 136 alunos de uma cidade de Ohio, os quais foram aleatoriamente divididos em dois grupos: um com 67 alunos (grupo 1) e o outro com os restantes 69 (grupo 2). Cada um dos grupos assistiu à mesma notícia relacionada com uma reunião e manifestação, naquela cidade, de uma organização racista, nascida nos finais do século XIX, nos EUA: Ku Klux Klan (KKK). A diferença para cada grupo consistiu no facto do ênfase final dado à notícia ser diferente em dois noticiários, cada um de um canal televisivo: enquanto um dos canais realçou esta reunião como sendo uma questão de liberdade de expressão (grupo 1), o outro sublinhou como uma questão de ameaça de ruptura da ordem pública (grupo 2). Após a exposição, a cada aluno foram apresentadas duas questões de forma a avaliar a sua abertura às reuniões e a estes discursos. A primeira consistiu em questionar o apoio ou a oposição, dos alunos, à permissão de manifestações públicas, naquela cidade, por membros do KKK; a segunda questionava o apoio ou oposição à permissão de discursos pelos membros do KKK. Para a mensuração da tolerância às reuniões e discursos recorreu-se a uma escala de 7 níveis, variando da categoria de "forte apoio" à categoria de "forte oposição". No contexto da análise de mediação causal, a variável de tratamento é a forma de abordagem dos meios de comunicação social, a variável mediadora é a opinião dos alunos relativamente à liberdade de expressão e à manutenção da ordem pública e a variável resposta é tolerância dos estudantes em relação ao KKK. Ou seja, os meios de comunicação social afectam a tolerância dos alunos ao KKK, podendo as suas reuniões e manifestações serem considerados pelos indivíduos como uma forma de liberdade de expressão ou de ordem pública (Imai, Keele e Yamamoto, 2010).

Atentando que a variável tratamento pode ser escrita como:

$$T_i = \begin{cases} 1, & se\ o\ aluno\ i\ pertence\ ao\ grupo\ 2 \\ 0, & se\ o\ aluno\ i\ pertence\ ao\ grupo\ 1' \end{cases}$$

ou seja, o grupo 2, o qual recebe a notícia que sublinha a reunião do KKK como uma questão de ordem pública, corresponde ao grupo de tratamento (ou simplesmente tratamento), enquanto o grupo 1 corresponde ao grupo de controlo; que existem três variáveis que funcionam como dependentes (mediadora e resposta), havendo a separação da mediadora em dois - importância em relação à liberdade de expressão e importância em relação à ordem pública -, e correspondendo a variável resposta à tolerância ao KKK, é apresentada a tabela 4.1. Na tabela apresenta-se a estimativa do efeito causal médio do tratamento (*ATE*) sobre cada variável que funciona como dependente, isto é, a estimativa do efeito médio do grupo 2 (relativamente ao grupo 1), sobre aquelas variáveis.

Tabela 4.1 - Resumo dos efeitos de tratamento médio estimados do estudo de Nelson, Clawson e Oxley (1997). Representação do efeito de tratamento médio estimado (*ATE*) da abordagem de ordem pública em relação à abordagem da liberdade de expressão e erro-padrão, para as três variáveis. Fonte: Imai, Keele e Yamamoto (2010).

Variáveis mediadoras e dependente	*ATE* (s.e.)
Importância da liberdade de expressão	-0.231 (0.239)
Importância da ordem pública	0.674 (0.303)
Tolerância ao KKK	-0.540 (0.340)

Considerando a definição do efeito causal médio do tratamento, os sinais das estimativas tabeladas estão de acordo com o esperado. O efeito da abordagem dos meios de comunicação social sobre a importância da liberdade de expressão origina uma estimativa negativa, visto o tratamento corresponder à notícia a favor da ordem pública. Isto constitui uma justificação para a estimativa positiva do efeito causal médio do tratamento (grupo 2) sobre a importância da ordem pública: o aluno ao ser exposto a uma notícia que realça uma possível ruptura de ordem pública influencia, de forma positiva, a sua opinião sobre a importância da existência de ordem pública. Sobre a tolerância ao KKK, a estimativa do *ATE* é negativa: se o tratamento correspondesse à atribuição da notícia abordando a liberdade de expressão, ao invés da ordem pública, possivelmente haveria uma maior tolerância, o que causaria uma estimativa positiva.

De notar que provavelmente ocorre uma violação da hipótese de ignorabilidade sequencial. Como referido, tratando-se de um estudo experimental, a atribuição de cada indivíduo ao grupo de tratamento ocorre de forma aleatória, o que satisfaz a condição de ignorabilidade do tratamento - expressão (4.31). O mesmo não se verifica com a ignorabilidade do mediador - expressão (4.32) -, pois sendo um estudo na área da Psicologia, não é possível randomizar as atitudes de um indivíduo. Consequentemente é provável que na relação mediador-resultado interfiram covariáveis não observadas (factores de confundimento), como por exemplo, a ideologia política de cada indivíduo, o que viola aquela hipótese fundamental de identificação de mecanismos causais (Imai, Keele e Yamamoto, 2010).

4.3.3. Identificação não paramétrica

Com a hipótese de ignorabilidade sequencial, é possível a identificação da distribuição de todos os resultados potenciais, $Y_i\big(t, M_i(t')\big)$. Esta distribuição depende de dois modelos: do modelo da variável resposta condicional à variável mediadora, ao tratamento e às covariáveis –$f(Y|M,T,X)$ - e do modelo do mediador condicional ao tratamento e às covariáveis - $f(M|\,T,X)$. Ou seja,

$$f\big(Y_i\big(t, M_i(t')\big)\big|X_i = x\big) = \int_{\mathcal{M}} f(Y_i|M_i = m, T_i = t, X_i = x)\, dF_{M_i}(m|T_i = t', X_i = x) \quad (4.33),$$

com $t, t' = 0{,}1$ e x pertencente ao suporte da distribuição de X_i, $\mathcal{X}$.

Com este resultado, através de valores observáveis das variáveis mediadora e resposta - por estarem sob um dado nível de tratamento observável -, é possível deduzir sobre as variáveis mediadoras e resposta submetidas ao nível de tratamento oposto, as quais não são observáveis. Ou seja, através de quantidades observadas (sob o nível de tratamento observável) são inferidas quantidades contrafactuais não observadas (sob o nível de tratamento oposto) (Imai, Tingley e Keele, 2010).

A expressão (4.33) representa a definição de identificação não paramétrica, apresentada por Imai, Keele e Tingley (2010), constituindo uma generalização do resultado obtido por Imai, Keele e Yamamoto (2010), os quais especificam o *ACME* e o *ADE*, respectivamente, como:

$$\bar{\delta}(t) = \iint E(Y_i \,|M_i = m, T_i = t, X_i = x)\left\{dF_{M_i \,|\, T_i=1, X_i=x}(m) - dF_{M_i \,|T_i=0, X_i=x}(m)\right\}dF_{X_i}(x) \quad (4.34)$$

$$\bar{\zeta}(t) = \iint \{E(Y_i|M_i = m, T_i = 1, X_i = x) - E(Y_i \,|M_i = m, T_i = 0, X_i = x)\}dF_{M_i \,|T_i=t, X_i=x}(m)dF_{X_i}(x) \quad (4.35),$$

com $t = 0{,}1$. A validade destes resultados mantém-se independentemente do modelo estatístico seguido pelas variáveis mediadora e resposta, sendo possível a extensão dos resultados a qualquer suporte da variável tratamento.

4.3.4. Algoritmo de estimação não paramétrica

Através da hipótese de identificação não paramétrica é possível estimar os efeitos causais de interesse envolvidos na análise de mediação, como o *ACME*, o *ADE* e o efeito total. Apenas é apresentado o algoritmo de estimação não paramétrico pois foi o utilizado na parte empírica da dissertação, tendo a vantagem de ser independente do modelo estatístico que as variáveis mediadora e a resposta seguem, o que simplifica a situação e não traz problemas. O método não paramétrico utilizado é de reamostragem de *Bootstrap*.

Após a obtenção das *J* amostras *bootstrap*, através da reamostragem com reposição, suponha-se que a quantidade de interesse a estimar é o *ACME*. Para cada reamostra, é necessário aplicar os seguintes passos:

Passo 1: Ajustar os modelos para as variáveis observáveis e mediadora, nomeadamente um modelo para a variável mediadora, $f(M_i|T_i, X_i)$, e um modelo para a variável resposta, $f(Y_i|M_i\,, T_i, X_i)$, usando os dados da amostra j, com $j = 1, 2, \ldots, J$. Os modelos serão designados por $f^{(j)}(M_i|T_i, X_i)$ e $f^{(j)}(Y_i|\, T_i, M_i\,, X_i)$, respectivamente.

Passo 2: Simular os valores potenciais que o mediador pode tomar, ${M_i}^{(jk)}(t)$, $k = 1, 2, \ldots K$. Estes valores simulados são obtidos de $f^{(j)}(M_i|\mathrm{t}, X_i)$, para cada $t = 0{,}1$ e cada $i = 1, 2, \ldots, n$ e $k = 1, 2, \ldots K$.

Passo 3: Simular os resultados potenciais tendo em conta os valores simulados do mediador. Os valores simulados dos resultados potenciais são obtidos ao retirar uma amostra de $f^{(j)}\left(Y_i\middle|t\,, {M_i}^{(jk)}(t')\,, X_i\right)$, ${Y_i}^{(jk)}\left(t, {M_i}^{(jk)}(t')\right)$, para cada $t = 0{,}1$ e cada $i = 1, 2, \ldots, n$ e $k = 1, 2, \ldots K$.

Passo 4: Calcular o efeito de mediação causal médio, como:

$$\bar{\delta}^{(j)}(t) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{{Y_i}^{(jk)}\left(t, {M_i}^{(jk)}(1)\right) - {Y_i}^{(jk)}\left(t, {M_i}^{(jk)}(0)\right)\right\} \quad (4.36).$$

Após a aplicação do algoritmo é possível calcular estatísticas de interesse, nomeadamente a estimativa pontual, o erro padrão e o IC, neste caso, do *ACME*, os quais são possíveis de obter da mediana, do desvio padrão e dos percentis da amostra de $\bar{\delta}^{(j)}(t)$, ou seja, através das *J* amostras *bootstrap* (Imai, Tingley e Keele, 2010).

4.3.5. Identificação dos efeitos no modelo de mediação simples sob a abordagem contrafactual

O modelo de mediação simples definido pelas equações de regressão linear (1.1), (1.2) e (1.3) pode ser escrito com a notação adoptada na abordagem contrafactual, como:

$$Y_i = i_1 + cT_i + \varepsilon_{i1} \quad (4.37)$$

$$Y_i = i_2 + c'T_i + bM_i + \varepsilon_{i2} \quad (4.38)$$

$$M_i = i_3 + aT_i + \varepsilon_{i3} \quad (4.39).$$

Como referido nas secções 4.2.5 e 4.2.6, o pressuposto de não interacção para o efeito médio de mediação causal (*ACME*) e o para o efeito directo médio (*ADE*) indica que estes efeitos não dependem do grupo de tratamento, isto é, assumindo que a variável tratamento só assume os valores 0 (grupo de controlo) ou 1 (grupo de tratamento), $\bar{\delta}(0) = \bar{\delta}(1) = \bar{\delta}$ e $\bar{\zeta}(0) = \bar{\zeta}(1) = \bar{\zeta}$. Nestas circunstâncias, o efeito total médio, $\bar{\tau}$, é dado como a soma destes dois efeitos: $\bar{\tau} = \bar{\delta} + \bar{\zeta}$.

Imai, Keele e Yamamoto (2010) provam que sob o pressuposto de não interacção do *ACME* e sob a hipótese de ignorabilidade sequencial definida na secção 4.3.1, a estimativa para o efeito indirecto como produto dos coeficientes ab é uma estimativa válida para o *ACME*. Ou seja, o *ACME* é identificado como $\bar{\delta}(0) = \bar{\delta}(1) = ab$. O mesmo se verifica para o *ADE*, isto é, $\bar{\zeta}(0) = \bar{\zeta}(1) = c'$, podendo o efeito total médio ser identificado por: $\bar{\tau} = ab + c' = c$.

No entanto, considerando o resultado de identificação não paramétrica, definido na secção 4.3.3, apenas a hipótese de ignorabilidade sequencial é necessária para identificar estes efeitos sobre o modelo de mediação simples (Imai, Keele e Tingley, 2010; Imai, Keele e Yamamoto, 2010).

4.3.6. "Proporção" mediada

Uma medida alternativa de avaliação da presença de mediação é a "proporção" mediada.[5]

No contexto da abordagem contrafactual, a "proporção" mediada representa a "magnitude dos efeitos médios de mediação causal em relação ao efeito total médio" (Imai, Tingley e Keele, 2010: 321), o que pode ser traduzido pelo quociente destas quantidades, ou seja:

$$\frac{\frac{\{\bar{\delta}(0) + \bar{\delta}(1)\}}{2}}{\tau} \quad (4.40)$$

(Imai, Tingley e Keele, 2010).

No contexto do modelo de mediação simples (secção 1.3) a "proporção" mediada corresponde à divisão do efeito indirecto pelo efeito total, sendo este último equivalente à soma do efeito directo com o efeito indirecto. Ou seja:

$$\frac{ab}{c' + ab} \quad (4.41)$$

(Ditlevsen, et al., 2005; Imai, Tingley e Keele, 2010).

Independentemente do contexto em que é definida, a "proporção" mediada reflecte quanto do efeito da variável tratamento sobre a variável resposta é explicada pela variável mediadora, razão pela qual é designada como "proportion explained" (Ditlevsen et al., 2005: 114). Esta medida não corresponde a uma proporção ou probabilidade, daí se escrever "proporção" (entre aspas), pois apesar de ser mencionada como tal, na realidade, trata-se de um rácio entre duas quantidades, podendo ser nula, negativa, ou positiva. No entanto, esta "proporção" só é lógica quando o numerador e o denominador têm o mesmo sinal, originando uma percentagem positiva. Especificamente ao se analisar a expressão (4.41), verifica-se que o quociente vai estar sempre definido entre 0 e 1: se c' (o efeito directo) for nulo o quociente dá a unidade; se os efeitos directo e indirecto forem ambos positivos ou ambos negativos, o quociente dá inferior a 1. Ou seja, esta medida, quando lógica, está sempre contida entre 0 e 1, mesmo não sendo uma proporção.

De forma a avaliar se esta medida é significativa, é necessário construir o respectivo intervalo de confiança, e verificar se contém o valor nulo. Este IC é calculado sobre aquele rácio.

5. Análise de Sensibilidade no contexto da abordagem Tradicional

Para a identificação não paramétrica dos efeitos, especialmente do *ACME*, é fundamental que a hipótese de ignorabilidade sequencial se verifique. Para tal, é necessário aplicar um conjunto de técnicas de análise de sensibilidade, sendo este recurso justificado pela impossibilidade de testar directamente, pelo conjunto de dados, a hipótese de ignorabilidade do mediador - expressão (4.32) -, ao invés da ignorabilidade do tratamento – expressão (4.31) -, a qual em estudos experimentais se verifica. A análise de sensibilidade permite investigar em que quantidade a hipótese de ignorabilidade do mediador deve ser violada, para que se registe uma alteração nas conclusões empíricas obtidas (Imai, Tingley e Keele, 2010). Ou analogamente, qual o desvio máximo permitido nesta hipótese para que as conclusões empíricas obtidas se mantenham (Imai, Keele e Yamamoto, 2010). No contexto da análise de mediação causal, a análise de sensibilidade revela-se globalmente importante, de forma a avaliar qualquer estudo de mediação.

Os autores Imai, Keele e Yamamoto (2010) desenvolveram técnicas de análise de sensibilidade baseadas no conjunto das equações (4.37), (4.38), (4.39), ou equivalentemente, (1.1), (1.2), (1.3), do modelo de mediação simples (Imai et al., 2017).

5.1. Análise de sensibilidade paramétrica com base na correlação entre os erros

O parâmetro de sensibilidade utilizado no contexto da análise de sensibilidade paramétrica consiste na correlação entre os erros da equação da variável resposta (4.38) e da equação da variável mediadora (4.39). Ou seja,

$$\rho = corr(\varepsilon_{i2}, \varepsilon_{i3}) \quad (5.1)$$

(Imai, Tingley e Keele, 2010; Imai, Keele e Yamamoto, 2010).

O parâmetro de sensibilidade ρ mede o grau de associação linear entre duas variáveis aleatórias, isto é, verifica se as duas variáveis se relacionam linearmente, podendo o seu valor variar entre -1 e 1 ($-1 < \rho < 1$). No caso do seu valor se aproximar de -1, indicará uma associação linear negativa entre as duas variáveis (neste caso, os dois erros), e quanto mais próximo de 1, uma associação linear positiva. À medida que o valor se aproximar de 0, menor será a associação linear, não existindo associação linear se tomar o valor 0.

A justificação do uso de ρ como parâmetro de sensibilidade remete para um ponto particular da secção 4.3.1: aquele que justifica a dificuldade em satisfazer a hipótese de ignorabilidade do mediador. Como referido, o condicionamento na hipótese de ignorabilidade do mediador é realizado nas covariáveis de pré-tratamento observáveis, o que possibilita a perturbação da relação causal mediador-resposta, pela existência de covariáveis com outro tipo de características (factores de confundimento): covariáveis de pós-tratamento (observáveis ou não) e covariáveis de pré-tratamento não mensuráveis. No entanto, a análise de sensibilidade desenvolvida apenas considera as covariáveis de pré-tratamento (Imai, Keele e Yamamoto, 2010). As covariáveis de pré-tratamento não mensuráveis são omitidas, sendo esta a justificação para a possível existência de uma associação entre os erros considerados no parâmetro de sensibilidade, visto os termos de erro incluírem as variáveis omitidas (Imai, Tingley e Keele, 2010).

Caso a hipótese de ignorabilidade sequencial seja satisfeita, o parâmetro ρ toma o valor nulo, o que assegura a inexistência de covariáveis de pré-tratamento omitidas. Caso contrário ρ toma valores não nulos (Imai, Tingley e Keele, 2010; Imai, Keele e Yamamoto, 2010).

Apesar da importância da hipótese de ignorabilidade sequencial na identificação dos efeitos de interesse, em particular do *ACME*, é possível definir e estimar consistentemente este efeito em função do valor do parâmetro de sensibilidade, desde que a ignorabilidade do tratamento se verifique (Imai, Tingley e Keele, 2010; Imai, Keele e Yamamoto, 2010). Fazendo variar o parâmetro de sensibilidade, obtêm-se estimativas para o *ACME*, isto é,

$$\bar{\delta}(0) = \bar{\delta}(1) = \frac{a\sigma_1}{\sigma_2}\left\{\tilde{\rho} - \rho\sqrt{\frac{1-\tilde{\rho}^2}{1-\rho^2}}\right\} \quad (5.2),$$

onde $\sigma_j, j = 1{,}2$ equivale ao desvio padrão dos erros, e $\tilde{\rho} \equiv Corr\,(\varepsilon_{i1}, \varepsilon_{i3})$ (Imai, Keele e Yamamoto, 2010).

O objectivo é verificar qual o grau de intensidade da correlação que provoca a inexistência deste efeito. Para tal, pode-se por um lado verificar se o intervalo de confiança para o *ACME* contém o valor zero, e por outro se ao variar ρ (partindo do principio que é nulo) obtêm-se diferenças no *ACME*, relativamente à sua estimativa quando a hipótese de ignorabilidade sequencial é satisfeita (Imai, Tingley e Keele, 2010). O efeito de mediação causal médio vem nulo caso $\tilde{\rho} = \rho$.

5.2. Análise de sensibilidade paramétrica com base nos coeficientes de determinação

Na análise de sensibilidade paramétrica baseada nos coeficientes de determinação, um conceito alternativo é dado ao parâmetro de sensibilidade ρ, com o objectivo de facilitar a sua interpretação (Imai, Tingley e Keele, 2010).

A omissão de covariáveis de pré-tratamento que influenciam a relação mediador-resultado (factores de confundimento de pré-tratamento), não mensuráveis, pode provocar a presença de correlação entre os erros, como referido na secção 5.1. Por conseguinte, o parâmetro ρ pode ser interpretado como a intensidade de um factor de confundimento (ou uma combinação linear de vários factores de confundimento) omitido(s) não observado(s), denotado(s) por U_i, para cada unidade amostral i. Ou seja, os erros ε_{i2} e ε_{i3} podem ser expressos como função de U_i:

$$\varepsilon_{i2} = \lambda_2 U_i + \varepsilon'_{i2} \quad (5.3)$$
$$\varepsilon_{i3} = \lambda_3 U_i + \varepsilon'_{i3} \quad (5.4),$$

onde $\lambda_j, j = 2{,}3$ representa o respectivo coeficiente desconhecido e $\varepsilon'_{ij} \perp\!\!\!\perp U_i, j = 2, 3$ (Imai, Tingley e Keele, 2010; Imai, Keele e Yamamoto, 2010).

O coeficiente de determinação é uma medida descritiva que mede ou avalia o ajustamento de um modelo de regressão, indicando em que quantidade o modelo é capaz de explicar os dados. O coeficiente de determinação varia entre 0 e 1, representando uma quantidade em proporção.

Com o ajustamento do modelo de mediador e de resultado, através das equações definidas no âmbito do conjunto de equações lineares, surgem os respectivos coeficientes de determinação, R_M^2 e R_Y^2 (Imai, Keele e Yamamoto, 2010). No entanto, não são estes os coeficientes de determinação de interesse no conceito do parâmetro de sensibilidade com base nos coeficientes de determinação. Ao invés, são

usados dois géneros de coeficientes de determinação, para o mediador e para a resposta: por um lado R_M^{2*} e R_Y^{2*}, e por outro $\tilde{R}_M^2$ e $\tilde{R}_Y^2$, com M indicando "Mediador" e Y o "Resultado". Os primeiros representam a proporção da variância que não é anteriormente explicada em cada uma das duas regressões do mediador e do resultado, sendo isto explicado pelo factor de confundimento de pré-tratamento omitido:

$$R_M^{2*} \equiv 1 - \frac{Var(\varepsilon'_{i2})}{Var(\varepsilon_{i2})} \quad (5.5)$$

$$R_Y^{2*} \equiv 1 - \frac{Var(\varepsilon'_{i3})}{Var(\varepsilon_{i3})} \quad (5.6).$$

A relação entre ρ e este par de coeficientes de determinação é dada por:

$$\rho^2 = R_M^{2*} R_Y^{2*} \quad (5.7).$$

Aplicando a raiz quadrada sobre este resultado obtém-se ρ:

$$\rho = \sqrt{R_M^{2*} R_Y^{2*}} = sgn(\lambda_2\lambda_3) R_M^* R_Y^* \quad (5.8),$$

onde $sgn(\lambda_2\lambda_3)$ representa a função sinal de $\lambda_2\lambda_3$, a qual é definida como:

$$sgn(\lambda_2\lambda_3) = \begin{cases} -1, & se\ \lambda_2\lambda_3 < 0 \\ 0, & se\ \lambda_2\lambda_3 = 0 \\ 1, & se\ \lambda_2\lambda_3 > 0 \end{cases} \quad (5.9).$$

Por outro lado, as expressões $\tilde{R}_M^2$ e $\tilde{R}_Y^2$ representam a proporção da variância original que é explicada em cada uma das duas regressões do mediador e do resultado, pelo factor de confundimento de pré-tratamento omitido:

$$\tilde{R}_M^2 \equiv \frac{Var(\varepsilon_{i2}) - Var(\varepsilon'_{i2})}{Var(M_i)} = (1 - R_M^2) R_M^{2*} \quad (5.10)$$

$$\tilde{R}_Y^2 \equiv \frac{Var(\varepsilon_{i3}) - Var(\varepsilon'_{i3})}{Var(Y_i)} = (1 - R_Y^2) R_Y^{2*} \quad (5.11).$$

A relação entre ρ e estas duas quantidades é dada por:

$$\rho^2 = \frac{\tilde{R}_M^2 \tilde{R}_Y^2}{\{(1 - R_M^2)(1 - R_Y^2)\}} \quad (5.12).$$

Aplicando a raiz quadrada sobre o resultado anterior obtém-se o parâmetro de sensibilidade:

$$\rho = \sqrt{\frac{\tilde{R}_M^2 \tilde{R}_Y^2}{\{(1 - R_M^2)(1 - R_Y^2)\}}} = \frac{sgn(\lambda_2\lambda_3)\tilde{R}_M \tilde{R}_Y}{\sqrt{(1 - R_M^2)(1 - R_Y^2)}} \quad (5.13)$$

(Imai, Tingley e Keele, 2010; Imai, Keele e Yamamoto, 2010).

6. Aplicação

6.1. Caracterização e objectivo do estudo base

O estudo não experimental subjacente à análise de mediação desenvolvida nesta dissertação abrangeu três zonas urbanas pertencentes à capital de Cabo Verde, cidade da Praia, nomeadamente, ao Platô, a uma parte de Vila Nova e a uma parte de Palmarejo, num âmbito de um projecto de investigação UPHI – STAT (PTDC/ATP-EUR/5074/2012). O bairro de Platô, ao qual foi dado o nome de zona formal, apresenta um planeamento urbano organizado, apresentando infra-estruturas, serviços básicos públicos, tendo água, esgotos, electricidade e recolha de lixo. A parte do bairro de Vila Nova foi designada por zona informal. Nesta zona, do ponto de vista da organização dos serviços e infra-estruturas apresenta uma maior carência. As estradas estão em mau estado, tal como os passeios ou calçadas. Na zona informal residem os grupos economicamente mais desfavorecidos. Por fim, a parte do bairro de Palmarejo constitui a zona de transição, apresentando uma combinação de aspectos das zonas formal e informal. O interesse nesta divisão refere-se às diferentes características das três zonas, quer em termos de desenvolvimento urbano, quer em termos sociais, económicos e culturais, pelo que o objectivo do estudo foi o de verificar de que forma as diversas características entre os bairros se reflectem na saúde dos seus habitantes (Gonçalves et al., 2015).

O estudo envolveu três fases e, no seguimento destas, diferentes dimensões de amostra. Na primeira foi aplicado um inquérito a uma amostra aleatória estratificada proporcional por zonas. Nesta primeira fase obteve-se um tamanho de amostra total de 1912 indivíduos. Considerando o objectivo deste estudo, o inquérito focou aspectos relacionados com a saúde e com o meio urbano, tendo sido subdividido em diferentes tópicos: caracterização sociodemográfica, o espaço onde vive, os equipamentos e bens que cada indivíduo tem na sua habitação, acessos a cuidados de saúde, estilos de vida quanto a hábitos tóxicos, alimentares e actividade física, à aquisição e compra de alimentos, a dados antropométricos autoreportados e reais e ainda à caracterização do padrão alimentar. Numa segunda fase, após o convite aos 1912 indivíduos, a amostra reduz-se para 599 indivíduos. Para estes, uma equipa de nutricionistas locais recolheu uma série de medidas, das quais fazem parte, altura, peso, massa muscular, gordura corporal, etc. Numa terceira fase, 118 indivíduos usaram pedómetros de forma a registar a quantidade de passos em cada dia, durante uma semana, havendo distinção entre dias úteis e não úteis (Gonçalves et al., 2015).

6.2. Análise exploratória de dados

No estudo de mediação é considerada a amostra inicial, isto é, o número total de indivíduos que responderam ao inquérito: 1912. O inquérito ao ser aplicado proporcionalmente a cada zona origina uma participação de 145, 1144 e 623 indivíduos na zona formal, de transição e informal, respectivamente. Atribuindo os indivíduos da zona de transição às zonas formal e informal verifica-se que 767 dos 1144 indivíduos da zona de transição são atribuídos à zona formal, totalizando esta zona 912 indivíduos; os restantes 377 habitantes da zona de transição são atribuídos à zona informal, totalizando esta zona 1000 indivíduos. Subjacente à atribuição está as características da zona de transição: a distribuição dos indivíduos pelas zonas formal e informal depende da sua residência em áreas mais ou menos carenciadas desta zona que combina áreas formais e informais. Assim, resulta uma variável binária que foi fundamental para aplicar a teoria da mediação aos casos apresentados posteriormente. No entanto, a análise exploratória de dados considera as três zonas, de acordo com o estudo base de Gonçalves et al. (2015).

Em qualquer zona foi maior a participação das mulheres – praticamente 60% do total de cada zona - comparativamente aos homens, apresentado a zona informal uma percentagem ligeiramente superior

relativamente às restantes zonas. Nas zonas de transição e informal habitam pessoas mais jovens, nomeadamente até aos 40 anos, mas a maioria apresenta idades inferiores aos 30 anos (aproximadamente 39.30% na zona de transição e 32.42% na zona informal). Contrariamente, a zona formal é constituída por indivíduos com mais idade, em particular por indivíduos com idades iguais ou superiores a 60 anos (35.21%). Este facto justifica a existência de uma significativa percentagem de reformados na zona formal (cerca de 30.34%), ocorrendo pouca diferenciação relativamente à percentagem de trabalhadores (cerca de 39.31%). A zona informal é aquela que tem maior percentagem de desempregados e menor percentagem de estudantes comparativamente às outras unidades urbanas, sendo a zona formal a que apresenta menor percentagem de desempregados e a zona de transição a que apresenta maior percentagem de estudantes. A razão pode ser devida às características das próprias zonas, como referido anteriormente. Em qualquer zona urbana predomina a percentagem de indivíduos que apresentam uma escolaridade média/alta. Contudo, na zona informal regista-se uma menor percentagem de indivíduos a concluírem a universidade. Relativamente à não escolaridade, a zona informal regista uma percentagem superior comparativamente às restantes zonas: 14.17%, contra os 3.55% e os 5.37%, da zona formal e de transição, respectivamente. Em relação ao estado civil, verifica-se em todas as zonas o mesmo padrão, existindo mais de 65% de indivíduos sem companheiro/a (solteiro, divorciado ou viúvo), apresentando a zona informal a percentagem superior (quase na ordem dos 75%). A maior percentagem de indivíduos casados ou em união de facto (cerca de 34.04%) regista-se na zona de transição. Apesar da maioria dos indivíduos, nas três zonas, não terem companheiro, em qualquer uma, é maior a percentagem de indivíduos que são pais comparativamente àqueles que não o são, apresentando a zona de transição a menor quantidade -cerca de 72% - e a zona informal a maior percentagem – cerca de 82%. Nesta última zona, a média de filhos é 3, observando-se dois indivíduos com 16 filhos, ficando as outras duas zonas por uma média de dois filhos. Em qualquer uma das zonas, a maioria dos seus indivíduos não reside no respectivo bairro desde a nascença, apresentando a zona de transição a maior percentagem de indivíduos "forasteiros" (cerca de 91.35%). Contudo, a zona informal contém a maior percentagem de indivíduos que dizem que residem lá desde o nascimento. Em relação à cidade da Praia, a maioria dos indivíduos da zona de transição também não reside lá desde a nascença. No entanto, a situação é diferente para as outras duas zonas: praticamente não existe diferença entre a percentagem de indivíduos a viverem ou não desde os seus nascimentos.

Quando questionados sobre a principal razão para residir no respectivo bairro, os indivíduos da zona formal e de transição optam pela tranquilidade como principal motivo (75.17% e 69.23%, respectivamente), enquanto os indivíduos da zona informal optam primeiramente pelas relações e habitações familiares (84.11%). Os motivos económicos e emprego no local não são principalmente apontados como razão, apresentando a zona informal a menor percentagem de inquiridos a referirem o emprego como razão, comparativamente às restantes zonas (cerca de 5.10%).

Em qualquer zona, de entre os nove problemas sociais apresentados, o desemprego é considerado como o mais grave pela maioria dos inquiridos, destacando-se na zona informal com um valor na ordem dos 92.12%. Além disso, esta zona apresenta as percentagens mais elevadas, quanto à classificação como grave, da pobreza/exclusão, assaltos/violência, custo de vida, droga e abandono e insegurança escolar. As zonas formal e de transição, para além do desemprego, também consideram como problemas mais graves o custo de vida, a droga e os assaltos/violência. A categoria "trânsito/acessibilidades" atinge uma maior percentagem na zona formal.

A maioria dos residentes das três zonas admite uma melhoria na qualidade de vida nos últimos 5 anos nos seus bairros, sendo que em qualquer zona, a maioria refere que é "bom" ou "satisfatório" residir na respectiva zona. Concretamente, a maioria dos inquiridos da zona formal admite que é "bom"

(61.54%), enquanto nas outras duas zonas já é menor a discrepância entre os que consideram que é "bom" ou "satisfatório". Nestas três zonas, a percentagem de residentes que considera que é "mau" habitar no respectivo bairro é reduzida, apresentando a zona informal a maior percentagem.

Com isto, questionou-se em que aspecto poderia ser melhorado o respectivo bairro, de forma a melhorar o bem-estar dos seus residentes. Os aspectos são a existência de espaços desportivos, manter os espaços públicos mais limpos, a existência de jardins e espaços verdes, a existência de espaços públicos, mais transportes públicos e melhores acessibilidades. No entanto, observa-se um padrão oposto quando se fala numa maior segurança: é evidente a necessidade de maior segurança nas três zonas, sobretudo na zona informal, que regista mais de 75% de respostas a favor. Outra necessidade nesta zona é também um melhor ambiente no geral (52.33%).

Foi igualmente questionado se estas pessoas praticavam actividade física no trabalho (se o trabalho obriga à prática de actividade física) e no tempo livre, e, no caso de resposta afirmativa, qual a sua intensidade. Verifica-se que, em qualquer zona, é superior o número de indivíduos que não pratica actividade física no trabalho (mais de 80%), apresentando, no entanto, a zona informal a maior percentagem de pessoas que o praticam (cerca de 17.17%). Relativamente à prática de actividade física no tempo livre, verifica-se o mesmo padrão: em qualquer zona, é maior a percentagem de indivíduos que não pratica actividade física, embora a discrepância entre o número de praticantes e não praticantes de actividade física seja menor relativamente à discrepância na prática de actividade física no trabalho. A zona formal apresenta a maior percentagem de praticantes no tempo livre em relação ao total da amostra, seguida da zona de transição e da zona informal (cerca de 40.56%, 39.02%, 30.39%, respectivamente). Relativamente à prática de exercício físico leve (andar a pé ou de bicicleta) durante pelo menos 10 minutos consecutivos nas suas deslocações, cerca de 60% dos indivíduos, em qualquer zona, é praticante. Em qualquer uma das zonas, a prática de exercício físico é realizada maioritariamente em espaços públicos. Em segundo lugar de frequência está o ginásio tanto na zona formal, como na zona de transição; para a zona informal surge o clube desportivo com cerca de 8.05% dos indivíduos. A piscina é o espaço físico menos usado com nenhum participante na zona formal e com uma percentagem quase nula na zona de transição e informal. Relacionado à prática de exercício físico está o meio de transporte utilizado para realizar deslocações. Em qualquer zona, verifica-se o mesmo padrão: o meio de transporte mais utilizado pelos inquiridos é o transporte público, o qual regista uma maior percentagem de uso na zona informal (cerca de 62.28%). Ao transporte público segue-se o táxi, vindo em ultimo a utilização de carro próprio, o qual atinge uma menor percentagem de uso na zona informal (cerca de 4.17%).

Esta breve síntese descritiva dos dados envolvidos no estudo base de Gonçalves et al. (2015) não abrange todas as suas variáveis nem todas aquelas fornecidas na base de dados desta dissertação, pois tornar-se-ia extenso. Não foram detalhadas as variáveis referentes às segundas e terceiras etapas do estudo (medição antropométrica e consumo alimentar e o uso de pedómetros, respectivamente). Como complemento, em anexo, são apresentadas duas tabelas, uma que resume algumas das variáveis qualitativas (anexo 1) e algumas das variáveis quantitativas (anexo 2) fornecidas, para além de alguns diagramas de boxplot para as últimas. As tabelas foram baseadas naquelas apresentadas no estudo de Gonçalves et al. (2015). Neste trabalho, escolheram-se algumas variáveis para ilustrar a aplicação da análise de mediação.

6.3. Análise de mediação

6.3.1. Síntese, objectivo e metodologia

De forma a aplicar a parte teórica, foram escolhidas algumas das variáveis disponibilizadas da base de dados do inquérito. Consideram-se três casos, no geral, os quais têm em comum a variável tratamento e resposta, mas diferentes variáveis mediadoras, sendo todas as variáveis binárias (já originalmente ou transformadas): a zona de residência dos indivíduos representa a variável tratamento, correspondendo o grupo de controlo à zona formal e o grupo de tratamento à zona informal; a variável resposta corresponde à prática de actividade física no tempo livre, a qual considera se um indivíduo realiza ou não exercício físico nesse período; a variável mediadora é especificada em cada caso.

Para a estimação dos efeitos envolvidos na análise de mediação recorreu-se à biblioteca *mediation* do programa estatístico *R*, especificamente à sua função *mediate*. Como referido, na abordagem contrafactual é possível identificar não parametricamente o efeito de mediação causal médio (*ACME*) e o efeito directo médio (*ADE*) assumindo a hipótese de ignorabilidade sequencial. Para tal, são necessários dois modelos: o modelo da variável resposta condicional à variável mediadora, ao tratamento e às covariáveis de pré-tratamento observáveis e o modelo da variável mediadora condicional ao tratamento e às mesmas covariáveis. A função *mediate* utiliza estes dois modelos como argumentos, os quais são ajustados separadamente. Como ambas as variáveis mediadora e resposta são binárias, em cada caso, os modelos ajustados foram modelos lineares generalizados. Em cada caso, foram ajustados três tipos de modelos lineares generalizados para o mediador e para a resposta, com base no número de covariáveis (nenhuma; idade; sexo, idade e tempo de residência na zona), para a totalidade dos indivíduos da amostra; e dois tipos de modelos tendo em conta a distinção por sexo dos indivíduos (modelo com nenhuma covariável ou com a covariável idade).

O método utilizado para estimar os efeitos envolvidos na análise de mediação causal, no contexto da abordagem contrafactual, foi o *Bootstrap* não paramétrico. O objectivo é verificar se a mediação é significativa nestes exemplos, comparando os respectivos modelos com e sem covariáveis. O anexo C apresenta o código em *R* utilizado na análise de mediação.

6.3.2. Dimensão da amostra

Foram apenas considerados os indivíduos que tinham valores para todas as variáveis (comando *complete.cases* no programa *R*) e não se recorreu à estimação de valores omissos. A covariável idade (em anos) foi introduzida na base de dados, sendo o seu cálculo possível através das variáveis correspondentes às datas de preenchimento do inquérito e de nascimento, em dias, meses e anos, estas sim presentes. Com a introdução da variável idade e com a opção apenas dos casos completos, a amostra diminui dos 1912 indivíduos iniciais para os 1275 indivíduos. Destes, 822 correspondem a elementos do sexo feminino e os restantes 453 a elementos do sexo masculino.

6.3.3. Caso 1

Como variável mediadora foi escolhida a variável correspondente às habilitações literárias, a qual considera, por um lado, os indivíduos sem escolaridade ou com escolaridade até ao ensino básico, e por outro, os indivíduos com escolaridade acima do ensino básico (ensino secundário e curso médio ou curso superior). Na base de dados original, esta variável assume cinco valores, pelo que os seus valores foram agrupados. A escolha desta variável como mediadora da relação entre a zona e a prática de exercício físico no tempo livre é justificada pelo facto de os indivíduos ao residirem em zonas de diferentes características (mais ou menos desenvolvidas), pode influenciar o seu nível de escolaridade, o que por sua vez influencia a sua prática de exercício físico no tempo livre.

6.3.3.1. Para a totalidade dos indivíduos

Os resultados, considerando a totalidade dos indivíduos, apresentam-se seguidamente tabelados.

Tabela 6.1 - Análise de mediação para a totalidade dos indivíduos, no caso 1. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção"[6] mediada média e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 1, para três tipos de modelo.

	Estimativas (Intervalo de confiança a 95%) (*n*=1275)					
	MODELO SEM COVARIÁVEIS		**MODELO COM UMA COVARIÁVEL (Idade)**		**MODELO COM VÁRIAS COVARIÁVEIS (Sexo, Idade, Tempo de residência na zona)**	
	Controlo	Tratamento	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0539 (-0.0752,-0.0400)	-0.0508 (-0.0715,-0.0300)	-0.0539 (-0.0738,-0.0400)	-0.0513 (-0.0696,-0.0300)	-0.0318 (-0.0452,-0.0200)	-0.0307 (-0.0436,-0.0200)
ADE	-0.0463 (-0.1018,0.0100)	-0.0432 (-0.0951,0.0100)	-0.0380 (-0.0945,0.0200)	-0.0354 (-0.0879,0.0200)	-0.0286 (-0.0786,0.0300)	-0.0275 (-0.0762,0.0300)
Média dos *ACME*	-0.0523 (-0.0737,-0.0400)		-0.0526 (-0.0711,-0.0300)		-0.0312 (-0.0441,-0.0200)	
Média dos *ADE*	-0.0448 (-0.0984,0.0100)		-0.0367 (-0.0912,0.0200)		-0.0280 (-0.0778,0.0300)	
Efeito Total Médio	-0.0971 (-0.1503,-0.0400)		-0.0894 (-0.1423,-0.0300)		-0.0592 (-0.1083,0.0000)	
"Proporção" Mediada Média	0.5388 (0.3158,1.2200)		0.5891 (0.3261,1.5400)		0.5271 (0.1755,2.9600)	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

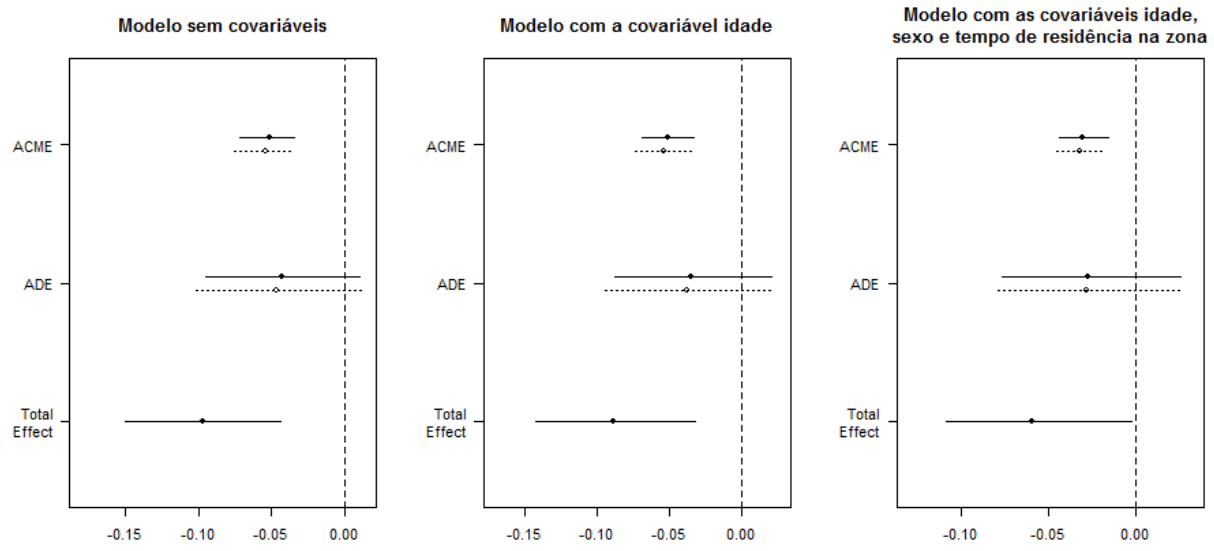


Figura 6.1 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, no caso 1. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 1, para os três tipos de modelo. As linhas e pontos a cheio são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

[6] A "proporção" mediada média não se refere a uma verdadeira proporção, daí o uso da expressão "proporção" entre aspas, em qualquer tabela ou legenda que a inclua.

Considerando a variável binária correspondente às habilitações literárias como variável mediadora, obtém-se uma conclusão geral válida para os três modelos estimados: o efeito de mediação causal médio e o efeito total médio estimados são estatisticamente significativos, visto o zero não estar contido nos respectivos intervalos de confiança. No entanto, pela tabela 6.1, verifica-se a inclusão de zero no intervalo de confiança correspondente ao efeito total médio, no modelo com três covariáveis. A conclusão de um efeito total médio estatisticamente significativo, no terceiro modelo, ilustra-se pela análise da figura 6.1, através da qual é possível considerar o zero incluído como um "zero negativo", pois o limite superior deste IC apesar de muito próximo deste valor, não o inclui. O efeito directo médio não é estatisticamente significativo em nenhum dos três modelos estimados.

Relativamente às estimativas do *ACME,* verifica-se uma diminuição nos seus valores, conforme a introdução de covariáveis, diferindo (pouco) entre as duas zonas. No modelo simples (sem covariáveis) regista-se uma diminuição ligeiramente superior, em média, na prática de exercício físico no tempo livre justificada pelas habilitações literárias - cerca de 5.39% - entre os indivíduos da zona formal (grupo de controlo), comparativamente à zona informal, a qual apresenta uma diminuição média de 5.08%. Como a variável mediadora tem suporte binário é possível especificar que, na zona formal, o facto de um indivíduo possuir escolaridade superior ao ensino básico provoca uma diminuição, em média, na prática de exercício físico no tempo livre de cerca de 5.39% superior relativamente aos indivíduos que possuem quanto muito o ensino básico. Na zona informal a discrepância é de cerca de 5.08%. As percentagens mantêm-se, praticamente, quando a covariável idade é introduzida no modelo, respectivamente, em 5.39% e 5.13%; a discrepância é superior quando as três covariáveis são consideradas: 3.18% e 3.07%, respectivamente, para a zona formal e informal.

Identicamente ao *ACME* observa-se um efeito total médio decrescente conforme a introdução de covariáveis. A variável zona causa uma diminuição total de cerca de 9.71% na realização de exercício físico no tempo livre, considerando o modelo simples. Este valor diminui para 8.94% e, mais acentuadamente para 5.92%, respectivamente, para os modelos com uma covariável (idade) e com três covariáveis (sexo, idade e tempo de residência na zona). Da diminuição total que se estima a zona causar na actividade física no tempo livre, cerca de 53.88% é explicada pelas habilitações literárias, considerando o modelo sem covariáveis. Com a introdução da covariável idade, a percentagem aumenta para 58.91% (diferença de cerca de 5.03%), diminuindo com o modelo de três variáveis, para cerca de 52.71% (diferença de cerca de 1.17%, em relação ao modelo simples). Ou seja, a percentagem do efeito total explicada pela variável mediadora, em média nas duas zonas, está acima dos 50%, em qualquer modelo estimado.

6.3.3.2. Estratificação por sexo

Os resultados, considerando a distinção por sexo, apresentam-se na seguinte tabela.

Tabela 6.2 - Análise de mediação para a totalidade dos indivíduos, com distinção de sexo, no caso 1. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção" mediada média e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 1.

MODELO SEM COVARIÁVEIS				
	Estimativas (Intervalo de confiança a 95%))			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0314 (-0.0572,-0.0100)	-0.0266 (-0.0532,-0.0100)	-0.0283 (-0.1342,-0.0100)	-0.0284 (-0.1322,-0.0100)
ADE	-0.0740 (-0.1230,0.0200)	-0.0692 (-0.1151,0.0200)	0.0131 (-0.1844,0.1200)	0.0129 (-0.1805,0.1200)
Média dos *ACME*	-0.0290 (-0.0552,-0.0100)		-0.0284 (-0.1323,-0.0100)	
Média dos *ADE*	-0.0716 (-0.1191,0.0200)		0.0130 (-0.1814,0.1200)	
Efeito Total Médio	-0.1007 (-0.1533,-0.0100)		-0.0154 (-0.2525,0.0600)	
"Proporção" Mediada Média	0.2884 (0.1026,1.8300)		-	
MODELO COM UMA COVARIÁVEL (Idade)				
	Estimativas (Intervalo de confiança a 95%)			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0279 (-0.0539,-0.0100)	-0.0241 (-0.0509,-0.0100)	-0.0311 (-0.1322,-0.0100)	-0.0316 (-0.1303,-0.0100)
ADE	-0.0647 (-0.1174,0.0300)	-0.0609 (-0.1106,0.0300)	0.0274 (-0.1773,0.1200)	0.0270 (-0.1750,0.1200)
Média dos *ACME*	-0.0260 (-0.0520,-0.0100)		-0.0314 (-0.1320,-0.0100)	
Média dos *ADE*	-0.0628 (-0.1135,0.0300)		0.0272 (-0.1760,0.1200)	
Efeito Total Médio	-0.0888 (-0.1417,0.0000)		-0.0042 (-0.2454,0.0700)	
"Proporção" Mediada Média	0.2926 (-0.0186,2.2500)		-	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

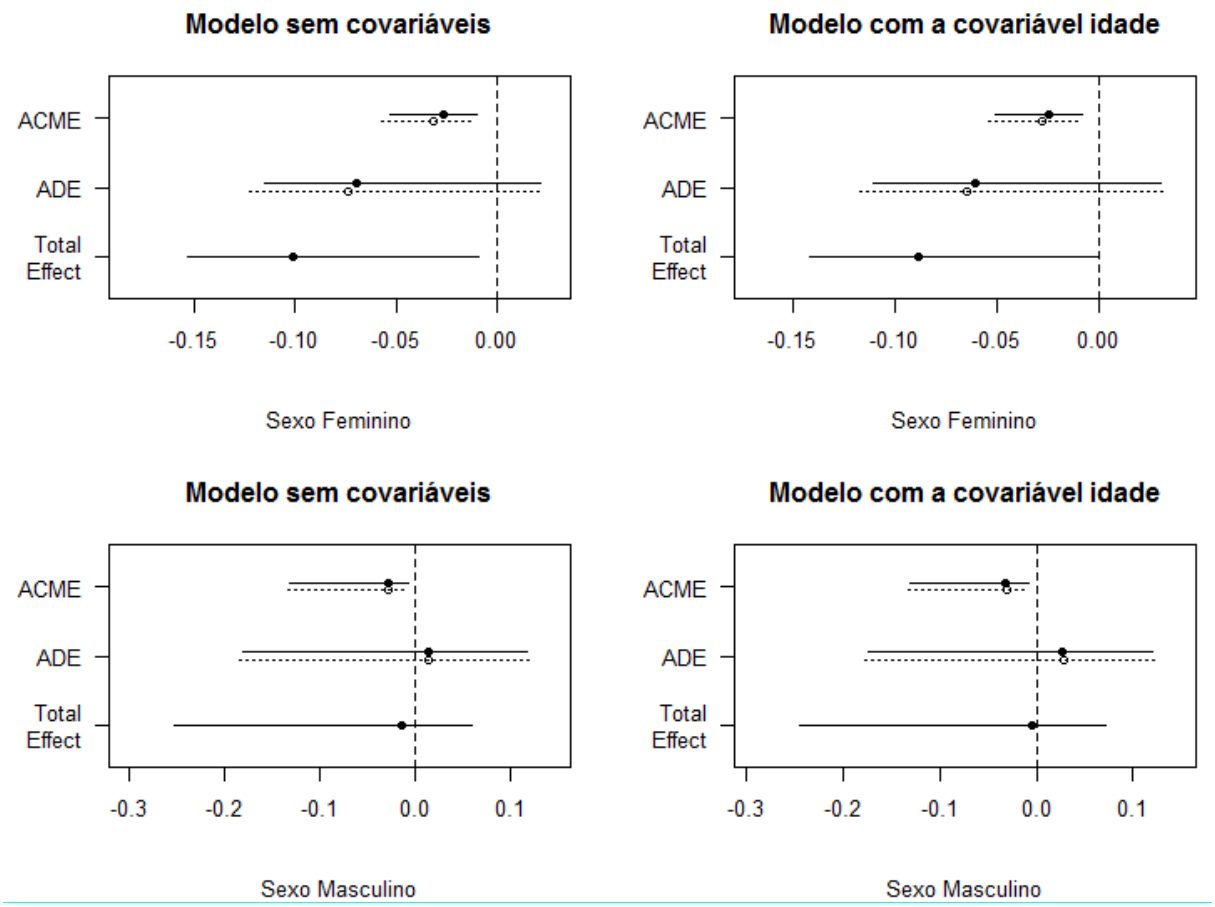


Figura 6.2 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, com distinção de sexo, no caso 1. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 1. As linhas e pontos a cheios são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

Estratificando os indivíduos por sexo, verifica-se a não inclusão de zero no IC estimado, pelos dois modelos (sem covariáveis e com a covariável idade), para o *ACME*, nos dois sexos. O oposto verifica-se em relação ao efeito directo médio. O efeito total médio é estatisticamente significativo quando se ajusta o modelo simples (sem covariáveis) entre o sexo feminino: torna-se estatisticamente não significativo quando se introduz a covariável idade no modelo e em qualquer modelo ajustado para o sexo masculino.

As estimativas do efeito de mediação causal médio, no geral, são superiores no sexo masculino comparativamente ao sexo feminino. Ajustando o modelo simples, regista-se uma maior diminuição, em média, na prática de exercício físico no tempo livre justificada pelas habilitações literárias - cerca de 3.14% - entre as mulheres da zona formal comparativamente às mulheres da zona informal, as quais apresentam uma diminuição média de 2.66%. Ajustando o modelo sem covariáveis para os homens, as estimativas (em módulo) são de 2.83% e 2.84%, respectivamente, na zona formal e informal. Equivalentemente referir que, entre o grupo feminino (masculino) da zona formal, o facto de uma mulher (homem) possuir habilitações superiores ao ensino básico provoca uma diminuição, em média, na prática de exercício físico no tempo livre de cerca de 3.14% (2.83%) superior àquelas (àqueles) que apenas possuem a escolaridade básica, é válido, visto a variável mediadora ser binária; entre as mulheres (homens) da zona informal a discrepância é de cerca 2.66% (2.84%). A introdução da covariável idade no modelo simples provoca certas diferenças. O padrão verificado no modelo simples para o sexo feminino mantém-se, contudo existe uma diminuição (em módulo) dos seus valores: entre

as mulheres da zona formal a diminuição, em média, na prática de exercício físico no tempo livre justificada pelas habilitações literárias, está na ordem dos 2.79%, enquanto o grupo de tratamento regista uma diminuição a rondar os 2.41%. Entre o sexo masculino, regista-se, ligeiramente, uma maior estimativa do *ACME* (em módulo) na zona informal comparativamente ao da zona formal e, também, um aumento (em módulo) nas estimativas deste efeito, entre os homens das duas zonas de estudo, comparativamente ao modelo simples. Ou seja, o efeito indirecto médio ronda os 3.16% entre os homens da zona informal e os cerca de 3.11% entre os do grupo de controlo. O único efeito total médio estatisticamente significativo (modelo sem covariáveis) indica que a variável zona causa uma diminuição de cerca de 10.07% na realização de exercício físico no tempo livre, para as mulheres. Assumindo o pressuposto de não interacção é possível a decomposição do efeito total médio no efeito de mediação causal médio e no efeito directo médio (estimativas resultantes da média das duas zonas). Da diminuição de 10.07%, em média, cerca de 2.90% é explicada pelas habilitações literárias das mulheres. Ou seja, da diminuição total que se estima a zona causar na actividade física no tempo livre das mulheres, cerca de 28.84% é explicada pela sua escolaridade, com o modelo simples.

De referir que ajustando o modelo para o grupo dos homens, a "proporção" mediada média é incoerente, visto as estimativas do *ACME* e do *ADE* médios ao apresentarem sinais opostos, compensam-se, resultando num efeito total médio inferior ao *ACME* médio. Consequentemente, o quociente entre o *ACME* médio e o efeito total médio resulta num número superior a 1. Como referido na secção 4.3.6, a incoerência não é devida ao facto de esta medida representar uma proporção, pois não o representa.

6.3.4. Caso 2

Como variável mediadora foi escolhida a variável correspondente à situação profissional, a qual considera, por um lado, os trabalhadores e por outro os não trabalhadores (desempregados, estudantes, reformados e domésticas). Na base de dados original, esta variável assume cinco valores, pelo que os seus valores foram agrupados. O facto de os indivíduos residirem em zonas de diferentes características (mais ou menos desenvolvidas) pode influenciar se se encontram activos a nível profissional ou não. Por sua vez, a situação profissional pode influenciar a prática de exercício físico no tempo livre: se os indivíduos estão empregados têm menos tempo para praticar desporto em relação aos desempregados, ou podem ter maior disponibilidade económica para pagar um ginásio, por exemplo.

6.3.4.1. Para a totalidade dos indivíduos

Os resultados, considerando a totalidade dos indivíduos, apresentam-se seguidamente tabelados.

Tabela 6.3 - Análise de mediação para a totalidade dos indivíduos, no caso 2. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção" mediada média e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 2, para três tipos de modelo.

	Estimativas (Intervalo de confiança a 95%) (*n*=1275)					
	MODELO SEM COVARIÁVEIS		**MODELO COM UMA COVARIÁVEL (Idade)**		**MODELO COM VÁRIAS COVARIÁVEIS (Sexo, Idade, Tempo de residência na zona)**	
	Controlo	Tratamento	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0171 (-0.0303,-0.0100)	-0.0156 (-0.0278,-0.0100)	-0.0139 (-0.0280,-0.0100)	-0.0129 (-0.0258,-0.0100)	-0.0026 (-0.0104,0.0000)	-0.0025 (-0.0100,0.0000)
ADE	-0.0831 (-0.1356,-0.0200)	-0.0817 (-0.1342,-0.0200)	-0.0748 (-0.1282,-0.0200)	-0.0737 (-0.1263,-0.0200)	-0.0568 (-0.1060,0.0000)	-0.0567 (-0.1059,0.0000)
Média dos *ACME*	-0.0164 (-0.0289,-0.0100)		-0.0134 (-0.0270,-0.0100)		-0.0025 (-0.0103,0.0000)	
Média dos *ADE*	-0.0824 (-0.1349,-0.0200)		-0.0742 (-0.1273,-0.0200)		-0.0567 (-0.1058,0.0000)	
Efeito Total Médio	-0.0988 (-0.1520,-0.0400)		-0.0876 (-0.1422,-0.0300)		-0.0593 (-0.1091,0.0000)	
"Proporção" Mediada Média	0.1656 (0.0622,0.4600)		0.1529 (0.0576,0.5500)		0.0428 (-0.1686,0.4000)	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

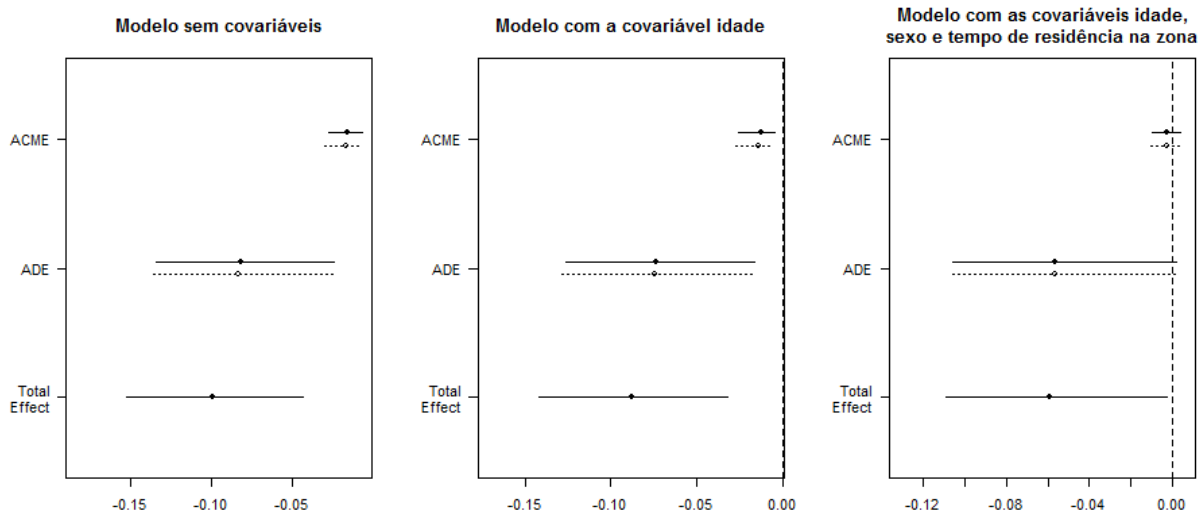


Figura 6.3 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, no caso 2. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 2, para os três tipos de modelo. As linhas e pontos a cheios são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

Relativamente à significância estatística, não se verificam diferenças tanto no modelo sem covariáveis como no modelo que considera a covariável idade - todos os efeitos e proporções mediadas são estatisticamente significativos. Considerando o modelo com as covariáveis sexo, idade e tempo de residência na zona não se verifica qualquer efeito estatisticamente significativo, de acordo com a tabela 6.3. No entanto, considerando o resumo gráfico 6.3, o IC do efeito total médio contém limites negativos, embora o limite superior esteja próximo de zero, tornando o efeito estatisticamente significativo.

Conforme a introdução de covariáveis no modelo simples, observa-se uma diminuição das estimativas do *ACME*, apresentando a zona formal um efeito ligeiramente superior em relação à zona informal. Ou seja, considerando qualquer modelo, regista-se uma diminuição ligeiramente superior, em média, na prática de exercício físico no tempo livre justificada pela situação profissional, entre os indivíduos da zona formal comparativamente aos residentes da zona informal. As diminuições percentuais correspondem, a 1.71% e 1.56% quando se considera o modelo simples (sem covariáveis); a 1.39% e 1.29% para o modelo com a covariável idade; e a 0.26% e 0.25% quando três covariáveis são consideradas (sexo, idade e tempo de residência na zona). Os últimos valores, como referido, não são estatisticamente significativos. Como o mediador considerado é binário, é possível referir equivalentemente que, por exemplo, o facto de um indivíduo se encontrar activo profissionalmente provoca uma diminuição média de cerca de 1.71% na prática de exercício físico no tempo livre, relativamente a um individuo que não o está, na zona formal (modelo sem covariáveis).

O mesmo padrão se observa para o *ADE*, isto é, considerando o modelo simples (sem covariáveis) e o modelo com a covariável idade, regista-se uma diminuição ligeiramente superior, em média, na prática de exercício físico no tempo livre, justificada pela própria zona (mantendo a situação profissional constante), entre os indivíduos do grupo de controlo comparativamente ao grupo de tratamento. Especificamente, as estimativas (em módulo) do efeito directo médio correspondem a 8.31% e 8.17% no modelo simples (sem covariáveis), reduzindo-se para 7.48% e 7.37% considerando o modelo com a covariável idade. No modelo com três covariáveis, as estimativas decrescem cerca de 2 pontos percentuais.

A análise anterior consistiu em comparar os efeitos (*ACME* e *ADE*) entre cada grupo de tratamento e de controlo. Uma comparação entre estes efeitos, dentro de cada zona, pode ser realizada. A diferença entre a diminuição na prática de exercício físico no tempo livre justificada pela própria zona (mantendo a situação profissional constante), comparativamente àquela justificada pela situação profissional, mantém-se praticamente constante com a introdução da covariável idade no modelo. Ou seja, considerando o modelo sem covariáveis, verifica-se uma menor diminuição na prática de exercício físico no tempo livre justificada pela situação profissional (cerca de 1.71%), do que pela própria zona (cerca de 8.31%), entre o grupo de controlo. A mesma conclusão se obtém para a zona informal. Considerando o modelo com a covariável idade, na zona formal, as estimativas do *ACME* e do *ADE* correspondem, respectivamente, a 1.39% e 7.48%; na zona informal as estimativas são de 1.29% e 7.37%, respectivamente.

Relativamente ao efeito causado pela variável zona na realização de exercício físico no tempo livre, verifica-se uma também diminuição na estimativa conforme a introdução de covariáveis no modelo. A variável zona causa uma diminuição de cerca de 9.88% na realização de exercício físico no tempo livre, considerando o modelo simples. A percentagem diminui para 8.76% e, mais acentuadamente para 5.93%, respectivamente, para os modelos com uma covariável e com três covariáveis. Da diminuição total que se estima a zona causar na actividade física no tempo livre, cerca de 16.56% é explicada pela situação profissional, usando o modelo sem covariáveis. Com a introdução da

covariável idade, a percentagem diminui para 15.29%, não sendo estatisticamente significativa para o modelo de três covariáveis.

6.3.4.2. Estratificação por sexo

Os resultados, considerando a distinção por sexo, apresentam-se seguidamente tabelados.

Tabela 6.4 - Análise de mediação para a totalidade dos indivíduos, com distinção de sexo, no caso 2. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção" mediada média e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 2.

MODELO SEM COVARIÁVEIS				
	Estimativas (Intervalo de confiança a 95%)			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0263 (-0.0442,-0.0100)	-0.0222 (-0.0404,0.0000)	0.0005 (-0.0149,0.0300)	0.0005 (-0.0155,0.030)
ADE	-0.0819 (-0.1327,0.0100)	-0.0778 (-0.1269,0.0100)	-0.0282 (-0.2557,0.0500)	-0.0282 (-0.2543,0.0500)
Média dos *ACME*	-0.0243 (-0.0419,0.0000)		0.0005 (-0.0156,0.0300)	
Média dos *ADE*	-0.0799 (-0.1291,0.0100)		-0.0282 (-0.2546,0.0500)	
Efeito Total Médio	-0.1041 (-0.1530,-0.0010)		-0.0277 (-0.2491,0.0700)	
"Proporção" Mediada Média	0.2330 (0.0326,1.2800)		-	
MODELO COM UMA COVARIÁVEL (Idade)				
	Estimativas (Intervalo de confiança a 95%)			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0253 (-0.0431,0.0000)	-0.0219 (-0.0397,0.0000)	0.0006 (-0.0148,0.0300)	0.0006 (-0.0152,0.0300)
ADE	-0.0697 (-0.1256,0.0200)	-0.0662 (-0.1211,0.0200)	-0.0219 (-0.2504,0.0600)	-0.0219 (-0.2475,0.0600)
Média dos *ACME*	-0.0236 (-0.0415,0.0000)		0.0006 (-0.0150,0.0300)	
Média dos *ADE*	-0.0680 (-0.1231,0.0200)		-0.0219 (-0.2505,0.0600)	
Efeito Total Médio	-0.0915 (-0.1444,0.0000)		-0.0213 (-0.2451,0.0700)	
"Proporção" Mediada Média	0.2576 (-0.3523,2.l400)		-	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

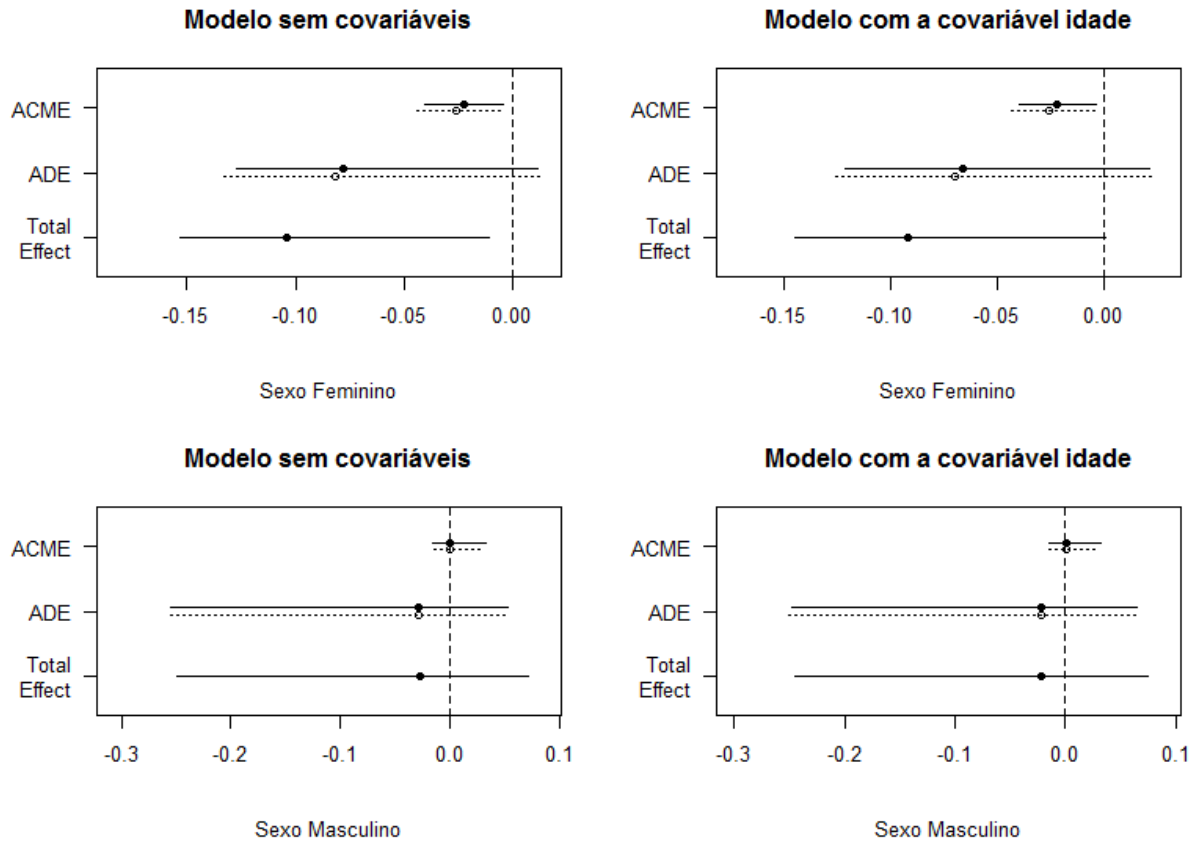


Figura 6.4 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, com distinção de sexo, no caso 2. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 2. As linhas e pontos a cheios são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

Estratificando por sexo, nenhuma estimativa surge como estatisticamente significativa para o sexo masculino, ao ajustar o modelo sem covariáveis e o modelo com a covariável idade. Entre o sexo feminino observam-se algumas diferenças. Em relação à significância estatística, apenas o *ADE*, entre as mulheres das duas zonas, não é estatisticamente significativo no modelo simples (sem covariáveis). Com a introdução da covariável idade, para além do *ADE*, também o efeito total médio e a proporção mediada média não o são. O *ACME* permanece estatisticamente significativo nos dois modelos ajustados. Embora a tabela 6.4 apresente a inclusão de zero nos intervalos de confiança estimados para o *ACME*, no modelo simples (para a zona informal) e no modelo com a covariável idade, pela figura 6.4, conclui-se que se trata de "zeros negativos", isto é, o limite superior dos intervalos de confiança foram aproximados a zero, não contendo, na realidade, esse valor.

Considerando o modelo simples, estima-se uma diminuição ligeiramente superior, em média, na prática de exercício físico no tempo livre justificada pela situação profissional - cerca de 2.63% -, entre o sexo feminino da zona formal comparativamente ao grupo de mulheres da zona informal (diminuição média de 2.22%). Equivalentemente é possível referir que o facto de uma mulher, residente na zona formal, estar activa profissionalmente provoca uma diminuição na prática de exercício físico no tempo livre de cerca de 2.63% superior àquelas que não estão; entre as mulheres da zona informal a discrepância é de cerca de 2.22%. O mesmo padrão se encontra nas estimativas obtidas com a introdução da covariável idade no modelo: o efeito de mediação causal médio é ronda os2.53% no grupo de controlo, e os 2.19% no grupo de tratamento.

O único efeito total médio estatisticamente significativo (modelo sem covariáveis) indica que a variável zona causa uma diminuição de cerca de 10.41% na realização de exercício físico no tempo livre, para as mulheres. Assumindo o pressuposto de não interacção é possível a decomposição do efeito total médio no efeito de mediação causal médio e no efeito directo médio (estimativas resultantes da média das duas zonas). Consequentemente, da diminuição de 10.41%, em média, cerca de 2.43% é explicada pela situação profissional das mulheres. Ou seja, da diminuição total que se estima a zona causar na actividade física no tempo livre das mulheres, cerca de 23.30% é explicada pela sua situação profissional (com o modelo simples).

De referir que, apesar de não significativa, a "proporção" mediada média resultante do ajustamento dos dois modelos para o sexo masculino é incoerente, pois o efeito total médio e a média dos *ACME* das duas zonas possuem sinais diferentes, o que provoca uma percentagem negativa.

6.3.5. Caso 3

Como variável mediadora foi escolhida a variável correspondente ao número de filhos, a qual considera, por um lado, os indivíduos com menos de três filhos, e por outro os indivíduos com 3 ou mais filhos. Na base de dados original a variável é discreta, pelo que os seus valores foram agrupados. A justificação da sua escolha como mediadora da relação zona - prática de exercício físico no tempo livre, está no facto de os indivíduos ao residirem em zonas de diferentes características (mais ou menos desenvolvidas) pode influenciar a quantidade de filhos que têm, o que por sua vez influencia a prática de desporto: se tiverem mais filhos, é provável que tenham menos tempo e eventualmente menos disponibilidade económica para a prática de actividade física no tempo livre.

6.3.5.1. Para a totalidade dos indivíduos

Os resultados, considerando a totalidade dos indivíduos, apresentam-se seguidamente tabelados.

Tabela 6.5 - Análise de mediação para a totalidade dos indivíduos, no caso 3. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção" mediada média e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 3, para três tipos de modelo.

	Estimativas (Intervalo de confiança a 95%) (*n*=1275)					
	MODELO SEM COVARIÁVEIS		**MODELO COM UMA COVARIÁVEL (Idade)**		**MODELO COM VÁRIAS COVARIÁVEIS (Sexo, Idade, Tempo de residência na zona)**	
	Controlo	Tratamento	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0177 (-0.0287,-0.0100)	-0.0161 (-0.0265,-0.0100)	-0.0168 (-0.0272,-0.0100)	-0.0154 (-0.0252,-0.0100)	-0.0039 (-0.0115,0.0000)	-0.0037 (-0.0109,0.0000)
ADE	-0.0843 (-0.1374,-0.0300)	-0.0827 (-0.1355,-0.0300)	-0.0750 (-0.1295,-0.0200)	-0.0736 (-0.1278,-0.0200)	-0.0554 (-0.1063,0.0000)	-0.0551 (-0.1058,0.0000)
Média dos *ACME*	-0.0169 (-0.0277,-0.0100)		-0.0161 (-0.0262,-0.0100)		-0.0038 (-0.0113,0.0000)	
Média dos *ADE*	-0.0835 (-0.1364,-0.0300)		-0.0743 (-0.1286,-0.0200)		-0.0552 (-0.1062,0.0000)	
Efeito Total Médio	-0.1004 (-0.1512,-0.0500)		-0.0904 (-0.1423,-0.0300)		-0.0591 (-0.1095,0.0000)	
"Proporção" Mediada Média	0.1685 (0.0631,0.4200)		0.1778 (0.0637,0.5100)		0.0646 (-0.0322,0.4600)	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

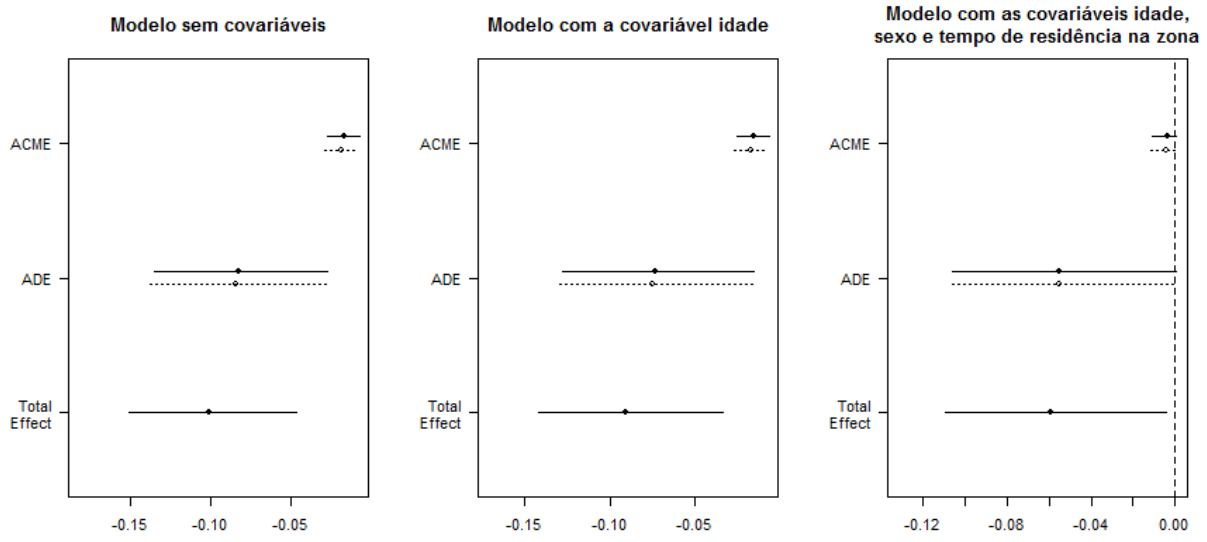


Figura 6.5 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, no caso 3. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para a totalidade dos indivíduos, no caso 3, para os três tipos de modelo. As linhas e pontos a cheios são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

As conclusões do presente caso assemelham-se ao caso 2. Qualquer efeito e proporção mediada resultantes do ajustamento do primeiro e segundo modelo (modelo simples e modelo com a covariável idade) são estatisticamente significativos. Com o ajustamento do modelo com as covariáveis sexo, idade e tempo de residência na zona, nenhum efeito mostra-se estatisticamente significativo, de acordo com a tabela 6.5. No entanto, considerando o resumo gráfico 6.5, o IC do efeito total médio contém limites negativos, embora o limite superior seja próximo de zero, o que torna o efeito estatisticamente significativo.

Conforme a introdução de covariáveis no modelo simples, as estimativas do *ACME* diminuem, não diferindo muito entre as duas zonas, apresentando a zona formal um efeito ligeiramente superior relativamente à zona informal. Ou seja, considerando qualquer modelo, verifica-se, entre os indivíduos da zona formal, uma diminuição superior, em média, na prática de exercício físico no tempo livre justificada pelo número de filhos, comparativamente à zona informal. As diminuições em percentagem correspondem, respectivamente, a 1.77% e 1.61% considerando o modelo simples (sem covariáveis); a 1.68% e 1.54% com o modelo com a covariável idade; e a 0.39% e 0.37% quando três covariáveis são consideradas (sexo, idade e tempo de residência na zona). Como o suporte da variável mediadora é binário, constata-se, por exemplo, que um indivíduo ao ter 3 ou mais filhos provoca uma diminuição, em média, na prática de exercício físico no tempo livre de 1.77% relativamente a um individuo com uma menor quantidade de filhos, na zona formal (modelo sem covariáveis).

O mesmo padrão é observado para o *ADE*, isto é, considerando o modelo simples (sem covariáveis) e o modelo com a covariável idade, verifica-se uma diminuição ligeiramente superior, em média, na prática de exercício físico no tempo livre, justificada pela própria zona (mantendo o número de filhos constante), entre os residentes da zona formal, relativamente à zona informal. O efeito directo médio, respectivamente, nas zonas formal e informal, assume (em módulo) os valores 8.43% e 8.27%, no modelo simples; e 7.50% e 7.36% no modelo com a covariável idade. Considerando o modelo com as três covariáveis, as estimativas decrescem cerca de 2 pontos percentuais.

A análise anterior compara os efeitos (*ACME* e *ADE*) entre cada grupo de tratamento e de controlo. Uma comparação entre os efeitos, dentro de cada zona, pode ser realizada. Considerando o modelo sem covariáveis, entre os indivíduos da zona formal, a diminuição na prática de exercício físico no tempo livre justificada pela própria zona ronda os 8.43%, consistindo num valor superior à percentagem explicada pelo número de filhos - cerca de 1.77%. Na zona informal verifica-se o mesmo, com igual diferença (cerca de 8.27% vs 1.61%). A diferença entre os dois efeitos, dentro de cada zona, mantém-se com a introdução de uma covariável (idade).

Por outro lado, verifica-se uma diminuição na estimativa do efeito total médio conforme a introdução de covariáveis no modelo. A variável zona causa uma diminuição de cerca de 10.04% na realização de exercício físico no tempo livre, por parte dos indivíduos, considerando o modelo simples. O valor percentual diminui para 9.04% e, mais acentuadamente para 5.91%, respectivamente, para os modelos com uma covariável e com três covariáveis. Da diminuição total que se estima a zona causar na actividade física no tempo livre, cerca de 16.85% é explicada pelo número de filhos, usando o modelo sem covariáveis. Com a introdução da covariável idade, o valor aumenta para 17.78%, não sendo estatisticamente significativo para o modelo de três covariáveis.

6.3.5.2. Estratificação por sexo

Os resultados, considerando a distinção por sexo, apresentam-se seguidamente tabelados.

Tabela 6.6 - Análise de mediação para a totalidade dos indivíduos, com distinção de sexo, no caso 3. Estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e média dos dois grupos, efeito total médio e "proporção" mediada média e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 3.

MODELO SEM COVARIÁVEIS				
	Estimativas (Intervalo de confiança a 95%)			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0096 (-0.0341,0.0000)	-0.0078 (-0.0030,0.0000)	-0.0038 (-0.0379,0.0100)	-0.0038 (-0.0382,0.0100)
ADE	-0.0949 (-0.1441,0.0000)	-0.0931 (-0.1392,0.0000)	-0.0223 (-0.2407,0.0800)	-0.0233 (-0.2407,0.0800)
Média dos *ACME*	-0.0087 (-0.0321,0.0000)		-0.0038 (-0.0381,0.0100)	
Média dos *ADE*	-0.0940 (-0.1418,0.0000)		-0.0223 (-0.2407,0.0800)	
Efeito Total Médio	-0.1027 (-0.1532,-0.0100)		-0.0261 (-0.2517,0.0700)	
"Proporção" Mediada Média	0.0845 (0.0076,0.8500)		0.1455 (-0.5515,1.3000)	
MODELO COM UMA COVARIÁVEL (Idade)				
	Estimativas (Intervalo de confiança a 95%)			
SEXO	**Feminino (n_F=822)**		**Masculino (n_M=453)**	
	Controlo	Tratamento	Controlo	Tratamento
ACME	-0.0086 (-0.0316,0.0000)	-0.0071 (-0.0287,0.0000)	-0.0042 (-0.0393,0.0100)	-0.0042 (-0.0391,0.0100)
ADE	-0.0840 (-0.1345,0.0100)	-0.0825 (-0.1319,0.0100)	-0.0128 (-0.2356,0.0800)	-0.0128 (-0.2348,0.0800)
Média dos *ACME*	-0.0079 (-0.0299,0.0000)		-0.0042 (-0.0388,0.0100)	
Média dos *ADE*	-0.0832 (-0.1334,0.0100)		-0.0128 (-0.2352,0.0800)	
Efeito Total Médio	-0.0911 (-0.1452,0.0000)		-0.0170 (-0.2442,0.0700)	
"Proporção" Mediada Média	0.0863 (-0.0348,1.0000)		0.2449 (-0.6186,1.2000)	

A representação gráfica das estimativas tabeladas consta na figura seguinte.

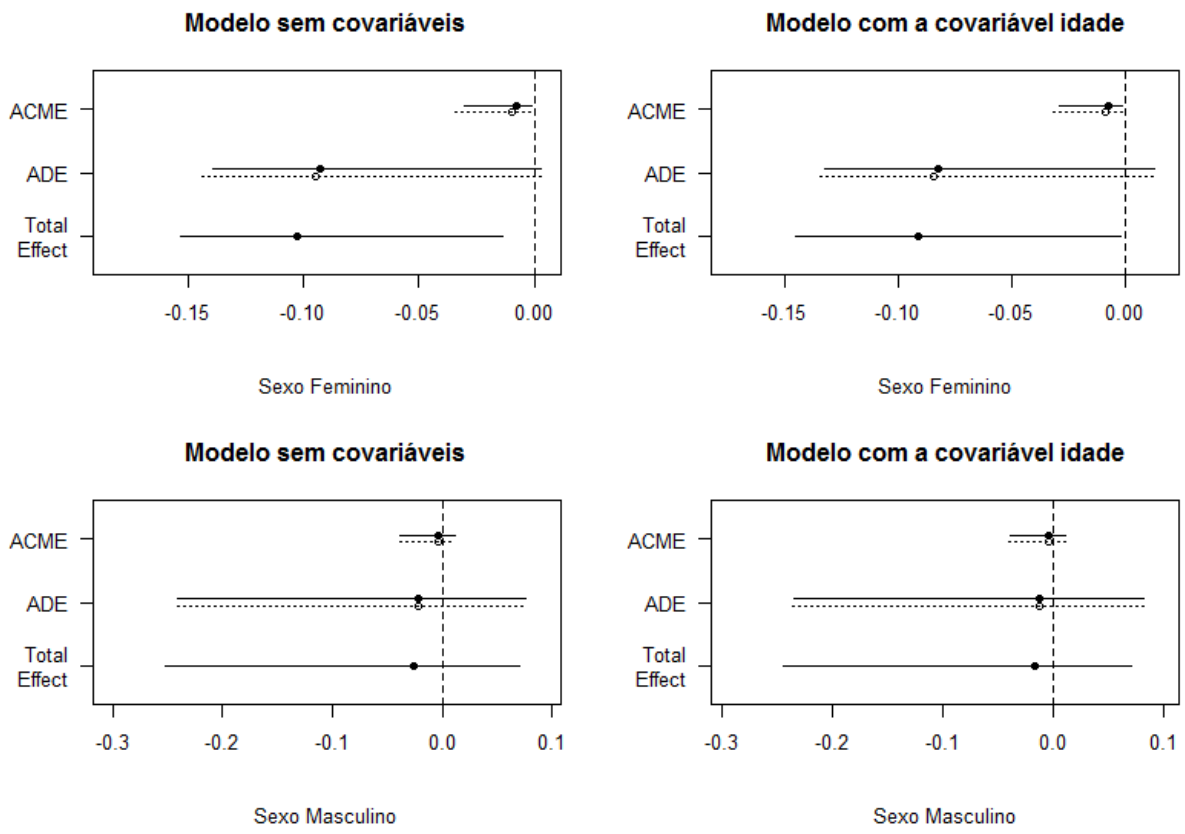


Figura 6.6 - Representação das estimativas dos efeitos e dos respectivos IC para a totalidade dos indivíduos, com distinção sexo, no caso 3. Representação das estimativas para os efeitos de mediação causal médio e directo médio para cada grupo e do efeito total médio e respectivos intervalos de confiança, para o sexo feminino e masculino, no caso 3. As linhas e pontos a cheios são referentes ao grupo de tratamento e as linhas a tracejado e pontos vazios são referentes ao grupo de controlo.

Estratificando por sexo, obtêm-se conclusões semelhantes ao caso 2, relativamente ao sexo masculino: nenhum efeito surge como estatisticamente significativo, em qualquer dos dois modelos ajustados. Entre o sexo feminino observam-se algumas diferenças. Relativamente à significância estatística entre as mulheres, verifica-se que o ajustamento do modelo simples (sem covariáveis) resulta num único efeito estatisticamente significativo, o efeito total médio. O seu valor indica que a variável zona causa uma diminuição de cerca de 10.27% na realização de exercício físico no tempo livre, para as mulheres. A introdução da covariável idade no modelo simples não altera estas conclusões. Apesar de zero estar contido no IC do efeito total médio no modelo com a covariável idade, constata-se, pela figura 6.6, que se trata de um "zero negativo", pois ambos os limites do intervalo de confiança são negativos.

7. Discussão

Nesta dissertação foram apresentadas abordagens para avaliar a mediação causal, existindo uma divisão em dois grupos: as abordagens baseadas nas estimativas das equações de regressão linear definidas no contexto do modelo de mediação simples e a abordagem contrafactual sob a identificação não paramétrica dos efeitos.

A primeira metodologia introduzida – abordagem dos passos causais ou método dos quatro passos -, é a metodologia mais utilizada, baseando-se em quatro passos, com o objectivo de analisar a significância estatística dos efeitos envolvidos (efeito total, efeito indirecto e efeito directo) no modelo de mediação. Com isto, surge a sua primeira desvantagem: o facto de não fornecer um cálculo directo do efeito mais interessante na mediação - o efeito indirecto - avaliando apenas a presença ou ausência de mediação (Newsom, 2015;[4]). Consequentemente não existe nenhum teste directo para avaliar a significância deste efeito[4], o qual se demonstra importante, ao invés de se testar somente a significância dos coeficientes nos quatro passos (Preacher e Hayes, 2004).

Por outro lado, de acordo com o conceito de mediação inconsistente, o facto da abordagem dos quatro passos exigir uma relação significativa entre a variável independente e a variável dependente (passo 1) reduz a potência em verificar a presença ou ausência de mediação, pois é possível existir realmente mediação sem este passo ser cumprido. Ou seja, esta abordagem apenas admite casos de mediação consistente - casos em que as estimativas dos efeitos directo e indirecto tenham sinais iguais e, portanto, a um valor da soma dos dois efeitos (o efeito total) não reduzido ou nulo, isto é, não significativo - (MacKinnon, Fairchild e Fritz, 2007; [2]). Consiste numa abordagem que pode levar a concluir ausência de mediação quando na verdade se verificam efeitos de mediação (erros de tipo II), daí apenas ser utilizada nos casos de mediação consistente (Newsom, 2015). Esta abordagem é recomendada perante amostras de dimensão grande (Preacher e Hayes, 2008).

Devido às desvantagens da metodologia dos passos causais, foram introduzidos testes que possibilitam avaliar directamente a significância do efeito indirecto, baseado na forma como se define este efeito: se como o produto ab (metodologia do produto de coeficientes) ou como a diferença $c - c'$ (metodologia da diferença de coeficientes). No entanto, os testes apresentam as suas limitações. Relativamente à primeira metodologia (teste de Sobel), Kenny[5] refere uma primeira crítica à forma como se obtém o erro padrão – expressão (2.3) e, mais usualmente, a expressão (2.4). Estas expressões são válidas no contexto do modelo de mediação simples definido na secção 1.3.1, isto é, num contexto de modelos de regressão linear, mas não noutros, como de regressão logística ou modelação de equações estruturais (*SEM*). Isto deve-se ao facto de estas expressões assumirem que as estimativas dos coeficientes que formam o efeito indirecto como produto de coeficientes, a e b, são independentes.

Por outro lado, estas duas últimas metodologias consideram que o rácio dos estimadores do efeito indirecto pelo respectivo erro padrão segue uma distribuição Normal, de forma a poder inferir sobre a significância estatística da presença de mediação. As metodologias do produto e da diferença de coeficientes assumem que o efeito indirecto segue uma distribuição Normal, no entanto, isto nem sempre se verifica - apenas em amostras de dimensão grande (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; Preacher e Hayes, 2008) -, fazendo com que estes testes de hipóteses apresentem uma baixa potência estatística (Preacher e Hayes, 2004; MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008;[5]). Como consequência, os intervalos de confiança referidos não são tão precisos, pois ao basearem-se numa distribuição simétrica, fá-los simétricos (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008).

Como solução deste problema, actualmente, são sugeridas outras metodologias para avaliar a significância estatística do efeito indirecto, como os métodos baseados na distribuição do produto de coeficientes e métodos de reamostragem, que têm implícitos intervalos de confiança assimétricos (MacKinnon, Fairchild e Fritz, 2007; MacKinnon, 2008; Preacher e Hayes, 2008). Dos últimos, destacam-se o método de *Monte Carlo* e o método *Bootstrap*, os quais, resumidamente, de forma a gerar uma distribuição empírica para o efeito indirecto, utilizam o conjunto de dados, tornando possível concluir sobre a significância estatística do efeito indirecto (MacKinnon, Fairchild e Fritz, 2007; Preacher e Hayes, 2008). Estas duas metodologias de reamostragem são recomendadas, pois sobrepõem-se em termos de potência estatística, tendo um maior controlo sobre o erro de Tipo I, relativamente quer à metodologia dos passos causais quer ao Teste de Sobel (Preacher e Hayes, 2008).

Como alternativa aos testes estatísticos com base na distribuição Normal são igualmente sugeridos o teste de significância conjunta de ab e o teste com base na distribuição do produto ab. No entanto, o primeiro não é muito utilizado, pois, além de pressupor que os coeficientes a e b não estão correlacionados, não é possível obter um intervalo de confiança.[5] Contudo é relatado como um bom teste, apresentando um "bom compromisso entre os erros de tipo I e tipo II" (MacKinnon, Fairchild e Fritz, 2007: p.601). O teste com base na distribuição do produto ab é aconselhado, pois visto ab não seguir sempre uma distribuição Normal, convém construir os respectivos IC através da distribuição do produto de duas variáveis. O método com base na distribuição do produto ab apresenta uma maior potência comparativamente ao método dos passos causais e às metodologias do produto e diferença de coeficientes (Preacher e Hayes, 2008).

Das metodologias apresentadas, são considerados como particularmente aconselháveis, para verificar a significância estatística do efeito indirecto (como ab), o método de *Bootstrap* e o método baseado na distribuição do produto (Preacher e Hayes, 2008).

Por último considerou-se a abordagem contrafactual, a qual define os efeitos envolvidos na mediação através da definição de resultados potenciais, tendo particular interesse o efeito de mediação causal médio, o efeito directo médio e o efeito total médio. Sob a hipótese de ignorabilidade sequencial é possível identificar os efeitos de forma não paramétrica, ou seja, obtêm-se expressões possíveis de aplicar a qualquer tipo de variáveis de tratamento, mediadoras e de resposta, admitindo-se qualquer tipo de relação entre as variáveis (lineares e não lineares), e não existindo uma dependência de método. Ou seja, não é necessário ter em consideração um qualquer modelo, sendo possível usar qualquer tipo de abordagem, paramétrica ou não paramétrica, qualquer que sejam os valores possíveis para a variável mediadora e resposta. Nesta dissertação foi desenvolvida a abordagem não paramétrica, nomeadamente através do *Bootstrap* não paramétrico. Esta consiste numa importante vantagem relativamente às abordagens que se baseiam no conjunto de equações lineares (1.1), (1.2) e (1.3). Ao se basearem nestas equações provoca, contrariamente ao relatado com o resultado de identificação não paramétrica, a não existência de uma "fórmula" geral para identificar os efeitos de interesse na mediação, visto ocorrer uma dependência relativamente ao suporte das variáveis mediadoras e resposta, ou seja, a cada modelo estatístico. Consequentemente, as metodologias que consideram o modelo de mediação simples definido na secção 1.3, são dificilmente estendidas a modelos não lineares. Outra vantagem associada à abordagem contrafactual apresentada no contexto da identificação não paramétrica é o facto de poder ser feita uma análise de sensibilidade, uma técnica que permite avaliar as conclusões obtidas e as hipóteses fundamentais (Imai, Tingley e Keele, 2010).

Referências bibliográficas

Abadie, A. (2005). "Causal Inference", *Encyclopedia of Social Measurement*, 1 (A-E): pp 259-266.

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M. T., & Keiding, N. (2005). "The Mediation Proportion: A Structural Equation Approach for Estimating the Proportion of Exposure Effect on Outcome Explained by an Intermediate Variable". *Epidemiology*, 16 (1): pp 114-120.

Esarey, J. (2015). "Causal Inference with Observational Data". In J. Bachner; K.V. Hill; B. Ginsberg (Eds.). *Analytics, Policy, and Governance*. New Haven: Yale University Press. pp 1-27.

Fauber, R., et al. (1990). "A Mediational Model of the Impact of Marital Conflict on Adolescent Adjustment in Intact and Divorced Families: The Role of Disrupted Parenting". *Child Development*, 61(4): pp 1112-1123.

Goldstein, N., D. (2016) "Epi Vignettes: Mediation frameworks and analysis". Disponivel em: http://www.goldsteinepi.com/blog/epivignettesmediationframeworksandanalysis (Consultado em 13/4/2017).

Gonçalves, L., et al. (2015). "Urban Planning and Health Inequities: Looking in a Small-Scale in a City of Cape Verde", *PLoS ONE*, 10(11): pp 1-27.

Holmbeck, G. N. (1997) "Toward Terminological, Conceptual, and Statistical Clarity in the Study of Mediators and Moderators: Examples From the Child-Clinical and Pediatric Psychology Literatures", *Journal of Consulting and Clinical Psychology*, 65 (4): pp 599-610.

Imai, K., Keele, L. Tingley, D., Yamamoto, T. (2017). "Causal Mediation Analysis Using R". In Imai *et al.* (Eds.). *Advances in Social Science Research Using R*. New York: Springer. pp 1-27. Disponível em: https://cran.r-project.org/web/packages/mediation/ (Consultado em 5/8/2017).

Imai, K., Keele, L., Yamamoto, T. (2010). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects", *Statistical Science*, 25(1): pp 51-71.

Imai, K., Tingley, D., Keele, L. (2010). "A General Approach to Causal Mediation Analysis", *Psychological Methods*, 15 (4): pp 309–334.

Imbens, G. W., Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Kenny, D.A. (s/d). "Estimating and Testing Mediation". Disponivel em: http://davidakenny.net/cm/MediationN.ppt (Consultado em 2/12/2016).

Kenny, D.A. (2016). "Mediation". Disponivel em: http://davidakenny.net/cm/mediate.htm (Consultado em 17/01/2017).

MacKinnon, D.P., Dwyer J. H. (1993). "Estimating Mediated Effects in Prevention Studies", *Eval Rev*, 17: pp 144-158.

MacKinnon, D. P. (2008). “Introduction to Statistical Mediation Analysis”, Taylor & Francis, New York.

MacKinnon, D. P., Fairchild, A. J., Fritz, M. S. (2007) “Mediation Analysis”, *Annual Review of Psychology,* 58: pp 593–614.

MacKinnon, D. P., Krull, J. K., Lockwood, C. M. (2000). “Equivalence of the Mediation, Confounding and Suppression Effect”, *Prevention Science,* 1(4): pp 173-181.

MacKinnon, D. P., Lockwood, C. M., Williams, J. (2004). “Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods”, *Multivariate Behavioral Research,* 39 (1): pp 99-128.

Morgan, K. L., Li, F. L. (2014). “SUTVA, Assignment Mechanism”. Disponível em: https://www2.stat.duke.edu/courses/Spring14/sta320.01/Class2.pdf (Consultado em 17/06/2017).

Nelson, T. E., Clawson, R. A., Oxley, Z. M. (1997). “Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance”, *American Political Science Review,* 91 (3): pp 567–583.

Newsom, J. (2015). “Testing Mediation with Regression Analysis”. Disponível em: web.pdx.edu/~newsomj/da2/ho_mediation.pdf (Consultado em 20/12/2016).

Pearl, J. (2001). “Direct and indirect effects”, *Proceedings of the 17th Conference on Uncertainty.* In J. S. Breese & D. Koller. *Artificial Intelligence (*Eds.). San Francisco, CA: Morgan Kaufmann. pp 411–420.

Preacher, K. J., Hayes, A. F. (2004). “SPSS and SAS procedures for estimating indirect effects in simple mediation models”, *Behavior Research Methods, Instruments, & Computers*, 36 (4): pp 717-731.

Preacher, K. J., Hayes, A. F. (2008). “Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models”, *Behavior Research Methods,* 40 (3): pp 879-891.

Qin, X. (2016). “An Introduction to Causal Mediation Analysis”. Disponível em: ww2.amstat.org/misc/webinarfiles/MHS3-09-2017.pdf (Consultado em 2/12/2016).

Rubin, D. B. (2005). “Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies”. Disponível em: http://www.stat.columbia.edu/~cook/qr33.pdf (Consultado em 23/07/2017)

Sales, A., C. (2016) “Review: mediation Package in R”, *Journal of Educational and Behavioral Statistics*, 42 (1): pp 69-84.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K. (2014). "mediation: R Package for Causal Mediation Analysis”: pp 1-40.

VanderWeele, T. J, Vansteelandt, S. (2009). "Conceptual issues concerning mediation, interventions and composition", *Statistics and Its Interface*, 2: pp 457–468.

Vieira, V. A. "Análise de Mediação e Moderação". Disponível em: https://marketinganpad.files.wordpress.com/2015/09/mediacao-e-moderacao-anpad-valter-afonso-vieira.pdf (Consultado em: 17/5/2017).

Yay, M. (2017). "The mediation analysis with the Sobel test and the percentile bootstrap", *International Journal of Management and Applied Science*, 3(2): pp 129 -131.

Wang, W. (2012). "Causal mediation analysis for non-linear models". Tese de Doutoramento em Filosofia. Department of Epidemiology and Biostatistics - Case Western Reserve University. 119 pp.

Wu, W. (2010). "Confounding, interaction, and mediation in multivariable/multivariate regression modeling". Disponível em: https://www.vumc.org/cqs/files/cqs/.../2010WilliamMay2010.pdf (Consultado: 8/5/2017).

Prado, P.H.M., Korelo, J.C., Silva, D.M.L. (2014). "Análise de Mediação, Moderação e Processos Condicionais", *Revista Brasileira de Marketing - ReMark*, 13(4): pp 4–24.

Anexo A: Análise das variáveis qualitativas fornecidas na base de dados

Tabela A.1 - Tabela descritiva das variáveis qualitativas. Distribuição dos indivíduos por zonas conforme os valores das variáveis qualitativas contidas na base de dados fornecida, relativa ao estudo de Gonçalves et al. (2015).

(*) Existência de observações omissas. A situação natural corresponde a uma zona formal, de transição e informal com 145, 1144 e 623 observações, respectivamente, pelo que, no caso de observações omissas, o valor da dimensão da(s) zona(s) em que tal acontece (e portanto o total de observações) é actualizado.

(**) Variável correspondente à 2º etapa do estudo de Gonçalves et al. (2015). Neste sentido, o número de observações correspondente ao total das três zonas reduziu-se de 1912 para 595.

Variáveis	**Número de indivíduos (Percentagem correspondente)**			
	Zona Formal	**Zona de Transição**	**Zona Informal**	**Total**
Zona	n_F = 145 (7.58%)	n_T = 1144 (59.83%)	n_I = 623 (32.58%)	n_{Total} = 1912 (100%)
Divisão da Zona em Formal e Informal	912 (47.70%)	-	1000 (52.30%)	1912 (100%)
Sexo				
Feminino	89 (61.38%)	729 (63.72%)	413 (66.29%)	1231 (64.38%)
Masculino	56 (38.62%)	415 (36.28%)	210 (33.71%)	681 (35.62%)
Idade [7] *	n_F^*= 142	n_T^* = 1122	n_I^* = 620	n_{Total}^* = 1884
18-29 anos	30 (21.13%)	441 (39.30%)	201 (32.42%)	672 (35.67%)
30-39 anos	25 (17.61%)	318 (28.34%)	144 (23.23%)	487 (25.85%)
40- 49 anos	21 (14.79%)	173 (15.42%)	93 (15.00%)	287 (15.23%)
50 – 59 anos	16 (11.27%)	114 (10.16%)	83 (13.39%)	213 (11.31%)
Mais de 60 anos	50 (35.21%)	76 (6.77%)	99 (15.97%)	225 (11.94%)
Habilitações Literárias *	n_F^*= 141	n_T^* = 1135	n_I^* = 621	n_{Total}^* = 1897
Sem Escolaridade/Nunca	5 (3.55%)	61 (5.37%)	88 (14.17%)	154 (8.12%)
Pré- escolar	7 (4.96%)	34 (3.00%)	41 (6.60%)	82 (4.32%)
Ensino Básico	31 (21.99%)	248 (21.85%)	191 (30.76%)	470 (24.78%)
Ensino Secundário e Curso médio	59 (41.84%)	418 (36.83%)	245 (39.45%)	722 (38.06%)
Curso superior	39 (27.66%)	374 (32.95%)	56 (9.02%)	469 (24.72%)
Profissão *	-	n_T^* = 1142	n_I^* = 622	n_{Total}^* = 1909
Trabalhador	57 (39.31%)	629 (55.08%)	240 (38.59%)	926 (48.51%)
Desempregado	9	200	179	388

[7] De notar que as variáveis Idade, IMC autoreportado e IMC real não são qualitativas mas sim contínuas, estando agrupadas em classes, daí a apresentação nesta tabela.

	(6.21%)	(17.51%)	(28.78%)	(20.32%)
Estudante	21 (14.48%)	204 (17.86%)	58 (9.32%)	283 (14.82%)
Reformado	44 (30.34%)	53 (4.55%)	51 (8.20%)	148 (7.75%)
Doméstica	14 (9.66%)	56 (4.90%)	94 (15.11%)	164 (8.59%)
Estado Civil *	-	n_T^* = 1134	n_l^* = 622	n_{Total}^* = 1901
Solteiro/Divorciado/Viúvo	100 (68.97%)	748 (65.96%)	462 (74.28%)	1310 (68.91%)
Casado/União facto	45 (31.03%)	386 (34.04%)	160 (25.72%)	591 (31.09%)
Filhos *	-	n_T^* = 1143	n_l^* = 622	n_{Total}^* = 1910
Não	30 (20.69%)	320 (28.00%)	110 (17.68%)	460 (24.08%)
Sim	115 (79.31%)	823 (72.00%)	512 (82.32%)	1450 (75.92%)
Tempo de residência na zona				
Não desde o nascimento	106 (73.10%)	1045 (91.35%)	390 (62.60%)	1541 (80.60%)
Desde o nascimento	38 (26.21%)	91 (7.95%)	228 (36.60%)	357 (18.67%)
Não sabe/Não responde	1 (0.69%)	8 (0.70%)	5 (0.80%)	14 (0.73%)
Tempo de residência na cidade da Praia				
Não desde o nascimento	69 (47.59%)	786 (68.71%)	299 (47.99%)	1154 (60.36%)
Desde o nascimento	72 (49.66%)	354 (30.94%)	319 (51.20%)	745 (38.96%)
Não sabe/Não responde	4 (2.76%)	4 (0.35%)	5 (0.80%)	13 (0.68%)
Se gosta de viver na respectiva zona *	-	n_T^* = 1128	n_l^* = 609	n_{Total}^* = 1882
Nada	5 (3.45%)	21 (1.86%)	13 (2.13%)	39 (2.07%)
Pouco	5 (3.45%)	36 (3.19%)	38 (6.24%)	79 (4.2%)
Indiferente	15 (10.34%)	165 (14.63%)	94 (15.44%)	274 (14.56%)
Bastante	46 (31.72%)	406 (35.99%)	235 (38.59%)	687 (36.50%)
Muito	74 (51.03%)	500 (44.33%)	229 (37.60%)	803 (42.67%)
Principal razão para residir na zona				
Tranquilidade	109 (75.17%)	792 (69.23%)	220 (35.31%)	1121 (58.63%)
Motivos económicos	32 (22.07%)	94 (8.22%)	72 (11.56%)	198 (10.36%)
Emprego no local	30 (20.69%)	156 (13.64%)	32 (5.14%)	218 (11.40%)
Relações/Habitação Familiar(es)	83 (57.24%)	495 (43.27%)	524 (84.11%)	1102 (57.64%)

Qual dos seguintes problemas considera grave na sua zona *				
Abandono/Insegurança escolar *	n_F^* = 144 9 (6.25%)	n_T^* = 1132 109 (9.63%)	n_I^* = 611 247 (40.43%)	n_{Total}^* = 1887 365 (19.34%)
Assaltos/Violência *	n_F^* = 143 34 (23.78%)	n_T^* = 1136 442 (38.91%)	n_I^* = 611 376 (61.54%)	n_{Total}^* = 1890 852 (45.08%)
Custo de vida *	n_F^* = 143 51 (35.66%)	n_T^* = 1137 437 (38.43%)	n_I^* = 609 352 (57.80%)	n_{Total}^* = 1889 840 (44.47%)
Desemprego *	n_F^* = 142 72 (50.70%)	n_T^* = 1136 662 (58.27%)	n_I^* = 609 561 (92.12%)	n_{Total}^* = 1887 1295 (68.63%)
Droga *	n_F^* = 143 43 (30.07%)	n_T^* = 1135 426 (37.53%)	n_I^* = 610 345 (56.56%)	n_{Total}^* = 1888 814 (43.11%)
Pobreza/Exclusão *	n_F^* = 141 28 (19.86%)	n_T^* = 1133 284 (25.07%)	n_I^* = 610 409 (67.05%)	n_{Total}^* = 1884 721 (38.27%)
Prostituição *	n_F^* = 141 13 (9.22%)	n_T^* = 1130 120 (10.62%)	n_I^* = 607 95 (15.65%)	n_{Total}^* = 1878 228 (12.14%)
Trânsito/Acessibilidades *	n_F^* = 140 25 (17.86%)	n_T^* = 1135 112 (9.87%)	n_I^* = 609 56 (9.20%)	n_{Total}^* = 1884 193 (10.24%)
Falta de civismo *	n_F^* = 142 27 (19.01%)	n_T^* = 1134 180 (15.87%)	n_I^* = 607 173 (28.50%)	n_{Total}^* = 1883 380 (20.18%)
Qualidade de vida nos últimos 5 anos na zona *	n_F^* = 142	n_T^* = 1133	n_I^* = 609	n_{Total}^* = 1884
Melhorou	71 (50.00%)	702 (61.96%)	278 (45.65%)	1051 (55.79%)
Piorou	12 (8.45%)	62 (5.47%)	93 (15.27%)	167 (8.86%)
Manteve-se igual	51 (35.92%)	249 (21.98%)	120 (19.70%)	491 (26.06%)
De uma maneira geral, viver na zona é: *	n_F^* = 143	n_T^* = 1117	n_I^* = 603	n_{Total}^* = 1863
Bom	88 (61.54%)	527 (47.18%)	258 (42.79%)	873 (46.86%)
Satisfatório	54 (37.76%)	561 (50.22%)	299 (49.59%)	914 (49.06%)
Mau	1 (0.70%)	13 (1.16%)	34 (5.64%)	48 (2.58%)
O que é necessário para se sentir melhor na zona				
Espaços desportivos	48 (33.10%)	479 (41.87%)	198 (31.78%)	725 (37.92%)
Jardins e espaços verdes	27 (18.62%)	443 (38.72%)	126 (20.22%)	596 (31.17%)
Maior segurança	90 (62.07%)	760 (66.43%)	475 (76.24%)	1325 (69.30%)
Espaços públicos	25 (17.24%)	296 (25.87%)	124 (19.90%)	445 (23.27%)

Limpeza nos espaços públicos	57 (39.31%)	477 (41.70%)	241 (38.68%)	775 (40.53%)
Mais transportes públicos	29 (29.00%)	180 (15.73%)	115 (18.46%)	324 (16.95%)
Melhores acessibilidades	28 (19.31%)	162 (14.16%)	126 (20.22%)	316 (16.53%)
Melhor ambiente no geral	59 (40.69%)	448 (39.16%)	326 (52.33%)	833 (43.57%)
Dieta prescrita *	n_F^* = 144	-	-	n_{Total}^* = 1911
Não	134 (93.06%)	1052 (91.96%)	553 (88.76%)	1739 (91.00%)
Sim	9 (6.25%)	92 (8.04%)	70 (11.24%)	171 (8.95%)
Actividade física no trabalho *	n_F^* = 140	n_T^* = 1143	-	n_{Total}^* = 1906
Não	117 (83.57%)	1031 (90.20%)	516 (82.83%)	1664 (87.30%)
Sim	23 (16.43%)	112 (9.80%)	107 (17.17%)	242 (12.70%)
Actividade física no tempo livre *	n_F^* = 143	n_T^* = 1143	n_I^* = 622	n_{Total}^* = 1908
Não	85 (59.44%)	697 (60.98%)	433 (69.61%)	1215 (63.68%)
Sim	58 (40.56%)	446 (39.02%)	189 (30.39%)	693 (36.38%)
Faz actividade física leve (pé ou bicicleta) durante pelo menos 10 minutos consecutivos *	n_F^* = 137	n_T^* = 1139	n_I^* = 622	n_{Total}^* = 1898
Não	57 (41.61%)	426 (37.40%)	244 (39.23%)	727 (38.30%)
Sim	79 (57.66%)	702 (61.63%)	375 (60.29%)	1156 (60.91%)
Numa semana normal, utiliza outros transportes				
Carro próprio	23 (15.86%)	186 (16.26%)	26 (4.17%)	235 (12.29%)
Transporte público	54 (37.24%)	660 (57.69%)	388 (62.28%)	1102 (57.64%)
Táxi	32 (22.07%)	214 (18.71%)	116 (18.62%)	362 (18.93%)
Onde pratica a actividade física *	n_F^* = 143	n_T^* = 1143	n_I^* = 621	n_{Total}^* = 1907
Ginásio *	9 (6.29%)	94 (8.22%)	19 (3.06%)	122 (6.40%)
Clube desportivo *	8 (5.59%)	43 (3.76%)	50 (8.05%)	101 (5.29%)
Piscina *	0	4 (0.35%)	3 (0.48%)	7 (0.37%)
Espaço público *	48 (33.57%)	335 (29.31%)	152 (24.48%)	535 (28.05%)
Outro e piscina *	6 (4.20%)	21 (1.84%)	10 (1.61%)	37 (1.94%)
IMC autoreportado [7]	n_F^* = 110	n_T^* = 901	n_I^* = 381	n_{Total}^* = 1392
Baixo peso	17	61	21	99

	(15.45%)	(6.77%)	(5.51%)	(7.11%)
Peso normal	53 (48.18%)	529 (58.71%)	191 (50.13%)	773 (55.53%)
Excesso de peso e Obesidade	40 (36.36%)	311 (34.52%)	169 (44.36%)	520 (37.36%)
Permissão para a realização de medidas				
Não	123 (84.83%)	861 (75.26%)	329 (52.81%)	1313 (68.67%)
Sim	22 (15.17%)	283 (24.74%)	294 (47.19%)	599 (31.33%)
IMC real [7] **	$n_F^{**} = 22$	$n_T^{**} = 281$	$n_I^{**} = 292$	$n_{Total}^{**} = 595$
Baixo peso	3 (13.64%)	23 (8.19%)	26 (8.90%)	52 (8.74%)
Peso normal	7 (31.82%)	109 (38.79%)	102 (34.93%)	218 (35.81%)
Excesso de peso e Obesidade	12 (54.55%)	149 (53.02%)	164 (56.16%)	325 (54.62%)

Anexo B: Análise das variáveis quantitativas fornecidas na base de dados

Tabela B.1 - Tabela descritiva das variáveis quantitativas. Indicação da mediana, do intervalo inter-quartil e do mínimo e máximo obtidos para as variáveis quantitativas contidas na base de dados fornecida, relativa ao estudo de Gonçalves et al. (2015).

	Média **Q_2= Mediana** **$(Q_1, Q_3) = (q_{0.25}, q_{0.75})$** **(Min, Max) = (Mínimo, Máximo)**			
VARIÁVEIS	**Zona Formal**	**Zona de Transição**	**Zona Informal**	**Total**
Tamanho do agregado	Média = 4 Q_2= 3 (Q_1, Q_3) = (2,4) (Min, Max) = (1,11)	Média = Q_2= 4 (Q_1, Q_3) = (3,5) (Min, Max) = (1,21)	Média = 5 Q_2 = 4 (Q_1, Q_3) = (3,6) (Min, Max) = (1,15)	Média = Q_2= 4 (Q_1, Q_3) = (3,5) (Min, Max) = (1,21)
Número de adultos (incluindo o próprio) no agregado familiar	Média = 3 Q_2= 2 (Q_1, Q_3) = (2,3) (Min, Max) = (1,8)	Média = Q_2= 3 (Q_1, Q_3) = (2,4) (Min, Max) = (1,18)	Média = Q_2= 3 (Q_1, Q_3) = (2,4) (Min, Max) = (1,11)	Média = Q_2= 3 (Q_1, Q_3) = (2,4) (Min, Max) = (1,18)
Número de filhos	Média = Q_2= 2 (Q_1, Q_3) = (1,4) (Min, Max) = (0,11)	Média = Q_2= 2 (Q_1, Q_3) = (0,3) (Min, Max) = (0,13)	Média = 3 Q_2 = 2 (Q_1, Q_3) = (1,4) (Min, Max) = (0,16)	Média = Q_2= 2 (Q_1, Q_3) = (1,3) (Min, Max) = (0,16)
Quantos quartos tem a casa onde vive	Média = Q_2= 3 (Q_1, Q_3) = (2,3) (Min, Max) = (1,10)	Média = Q_2= 3 (Q_1, Q_3) = (2,3) (Min, Max) = (1,12)	Média = Q_2= 2 (Q_1, Q_3) = (1,3) (Min, Max) = (1,18)	Média = Q_2= 3 (Q_1, Q_3) = (2,3) (Min, Max) = (1,18)
Número de horas sentado por dia, em média	Média = 4.829 Q_2=5 (Q_1, Q_3) = (3,6) (Min, Max) = (0.5,12)	Média = 5.108 Q_2=5 (Q_1, Q_3) = (3,7) (Min, Max) = (0,14)	Média = 4.501 Q_2=4 (Q_1, Q_3) = (3,6) (Min, Max) = (0,16)	Média = 4.895 Q_2=5 (Q_1, Q_3) = (3,6) (Min, Max) = (0,16)
Número de horas de sono por dia, em média	Média = 7.379 Q_2 = 8 (Q_1, Q_3) = (6,8) (Min, Max) = (2,12)	Média = 7.518 Q_2 = 8 (Q_1, Q_3) = (7,8) (Min, Max) = (1,19)	Média = 7.425 Q_2 = 8 (Q_1, Q_3) = (6,8) (Min, Max) = (2,12)	Média = 7.477 Q_2 = 8 (Q_1, Q_3) = (7,8) (Min, Max) = (1,19)
Altura real (m)	Média = 1.60 Q_2 = 1.59 (Q_1, Q_3) = (1.56, 1.64) (Min, Max) = (1.5,1.72)	Média = 1.63 Q_2 = 1.61 (Q_1, Q_3) = (1.56, 1.68) (Min, Max) = (1.44,1.87)	Média = 1.63 Q_2 = 1.61 (Q_1, Q_3) = (1.57, 1.69) (Min, Max) = (1.45,1.92)	Média = 1.63 Q_2 = 1.61 (Q_1, Q_3) = (1.56, 1.68) (Min, Max) = (1.44,1.92)
Peso sem roupa nem sapatos (kg)	Média = 69.15 Q_2 = 70 (Q_1, Q_3) = (60, 75) (Min, Max) = (40,100)	Média = 68.09 Q_2 = 68 (Q_1, Q_3) = (60, 76) (Min, Max) = (40,138)	Média = 69.61 Q_2 = 70 (Q_1, Q_3) = (60, 78) (Min, Max) = (40,125)	Média = 68.61 Q_2 = 68 (Q_1, Q_3) = (60, 76) (Min, Max) = (40,138)
Peso desejado (kg)	Média = 65.85 Q_2 = 65 (Q_1, Q_3) = (60, 74) (Min, Max) = (44,93)	Média = 65.23 Q_2 = 65 (Q_1, Q_3) = (59, 70) (Min, Max) = (45,105)	Média = 66.02 Q_2 = 65 (Q_1, Q_3) = (60, 71) (Min, Max) = (49,100)	Média = 65.53 Q_2 = 65 (Q_1, Q_3) = (60, 70) (Min, Max) = (44,105)

Peso real (kg)	Média = 70.16 Q_2 = 68.25 (Q_1, Q_3) = (57.15, 81.68) (Min, Max) = (46.80,99.10)	Média = 68.28 Q_2 = 67.15 (Q_1, Q_3) = (57.80, 78.08) (Min, Max) =(38.20,116.80)	Média = 70.28 Q_2 = 68.60 (Q_1, Q_3) = (60.50, 78.73) (Min, Max) =(38.00,128.20)	Média = 69.33 Q_2 = 67.75 (Q_1, Q_3) = (58.68, 78.55) (Min, Max) =(38.00,128.20)
IMC autodeclarado (kg/m^2)	Média = 24.61 Q_2 = 24.39 (Q_1, Q_3) = (21.80, 26.67) (Min, Max) = (16.41,33.20)	Média = 24.26 Q_2 = 23.94 (Q_1, Q_3) = (21.68, 26.37) (Min, Max) = (14.53,41.78)	Média = 25.30 Q_2 = 24.65 (Q_1, Q_3) = (21.92, 27.70) (Min, Max) = (14.69,47.63)	Média = 24.57 Q_2 = 24.22 (Q_1, Q_3) = (21.72, 26.67) (Min, Max) = (14.53,47.63)
IMC real (kg/m^2)	Média = 27.52 Q_2 = 26.45 (Q_1, Q_3) = (22.01, 31.45) (Min, Max) = (17.82,44.04)	Média = 25.85 Q_2 = 25.49 (Q_1, Q_3) = (21.60, 29.30) (Min, Max) = (14.01,41.25)	Média = 26.56 Q_2 = 25.92 (Q_1, Q_3) = (22.34, 30.17) (Min, Max) = (15.42,48.25)	Média = 26.26 Q_2 = 25.81 (Q_1, Q_3) = (21.95, 29.87) (Min, Max) = (14.01,48.25)

Esta tabela pode ser resumida pelos seguintes gráficos boxplot:

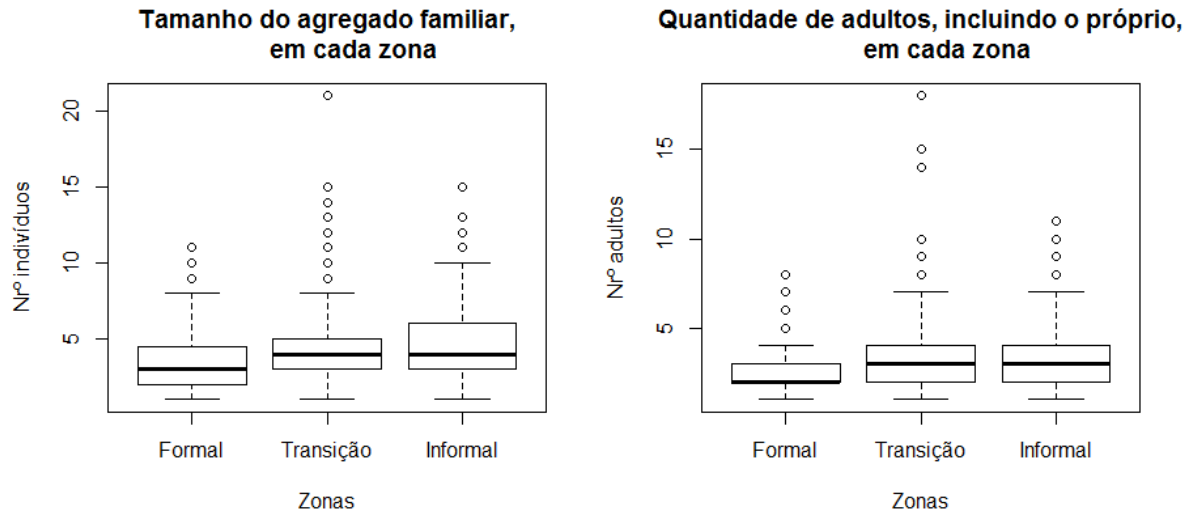


Figura B.1 – Boxplot correspondente ao tamanho do agregado familiar e ao número de adultos que o constituem. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo das variáveis quantitativas correspondentes ao tamanho do agregado familiar do inquirido e ao número de adultos, inclusive o inquirido, que o constituem, por zona.

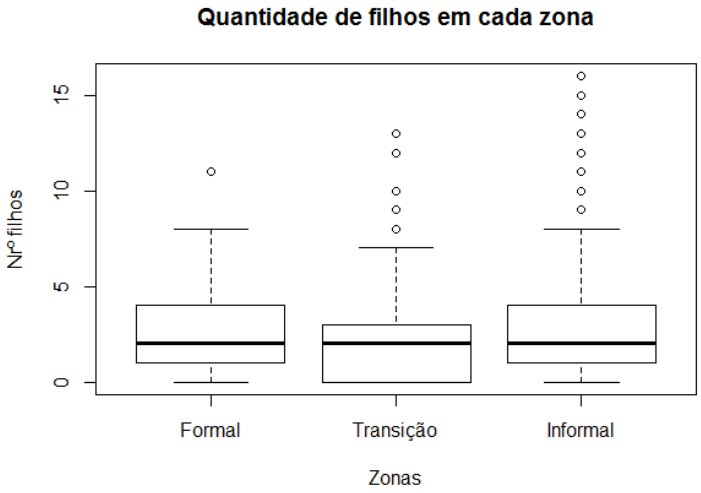


Figura B.2 – Boxplot correspondente ao número de filhos. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo da variável quantitativa correspondente ao número de filhos de cada individuo, por zona.

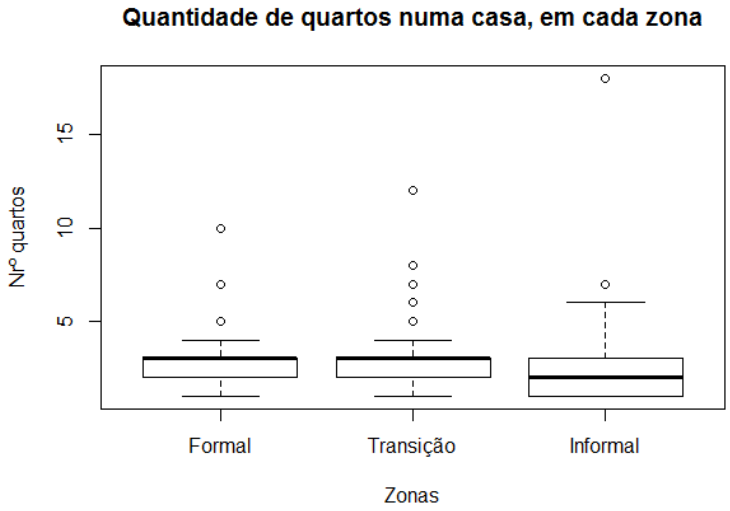


Figura B.3 - Boxplot correspondente ao número de quartos existente na residência do indivíduo. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo da variável quantitativa correspondente ao número de quartos existente na residência de cada indivíduo, por zona.

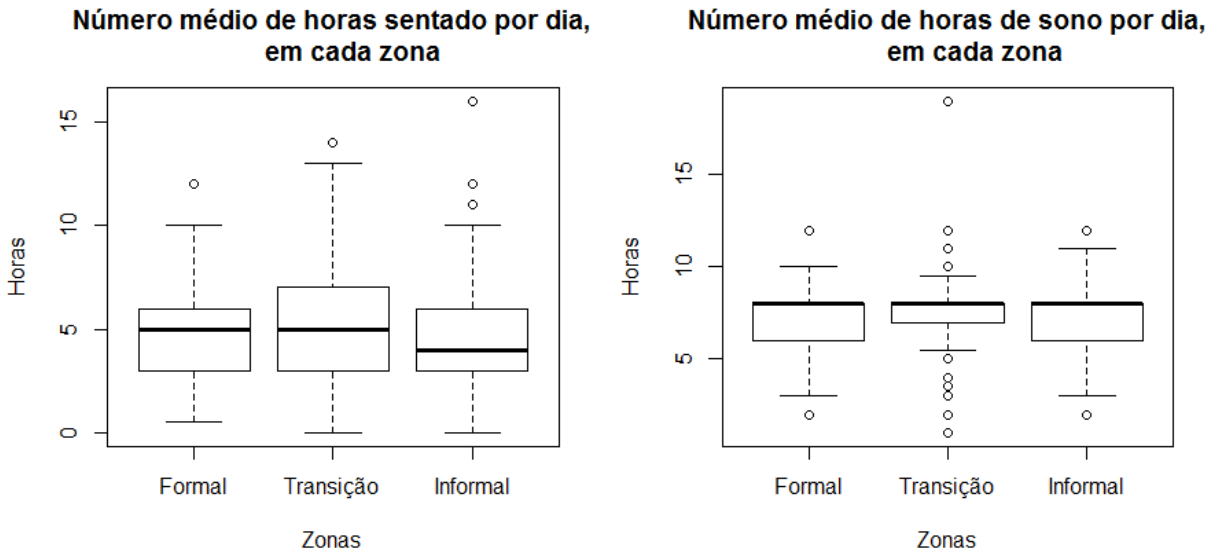


Figura B.4 - Boxplot correspondente ao número de horas sentado e de sono por dia. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo das variáveis quantitativas correspondentes ao número médio de horas que o inquirido passa sentado e dorme por dia, por zona.

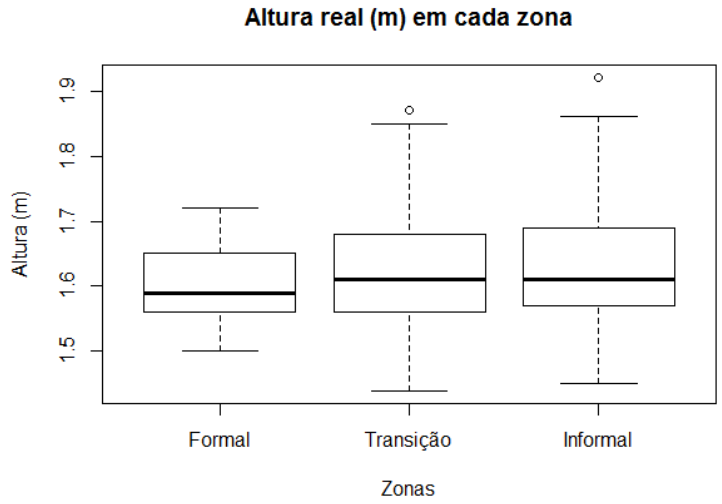


Figura B.5 - Boxplot correspondente à altura real. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo da variável quantitativa correspondente à altura (m) real do inquirido, por zona.

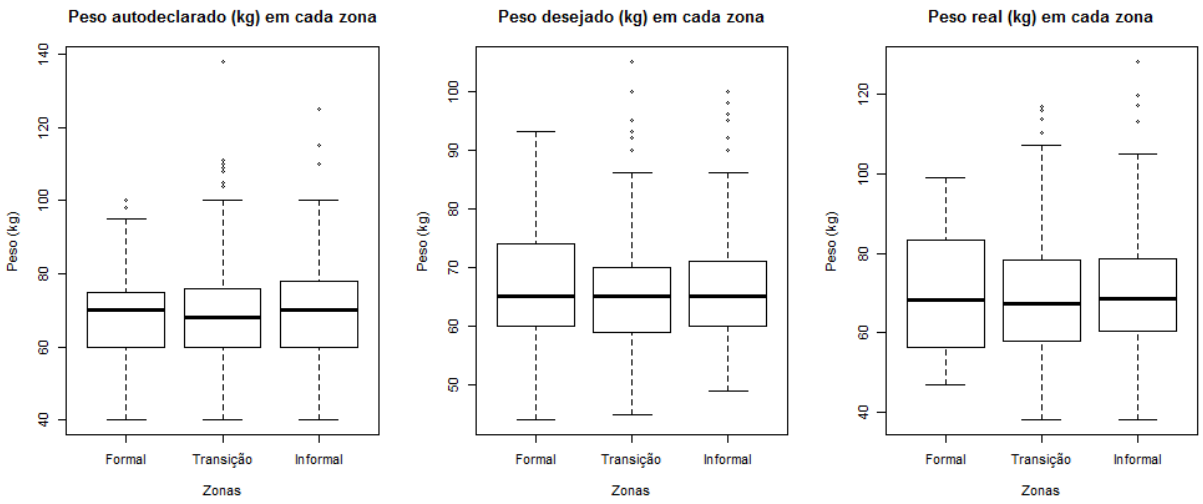


Figura B.6 - Boxplot correspondente ao peso autodeclarado, desejado e real. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo das variáveis quantitativas correspondentes ao peso (kg) habitual, desejado e real, sem roupa nem sapatos, do inquirido, por zona.

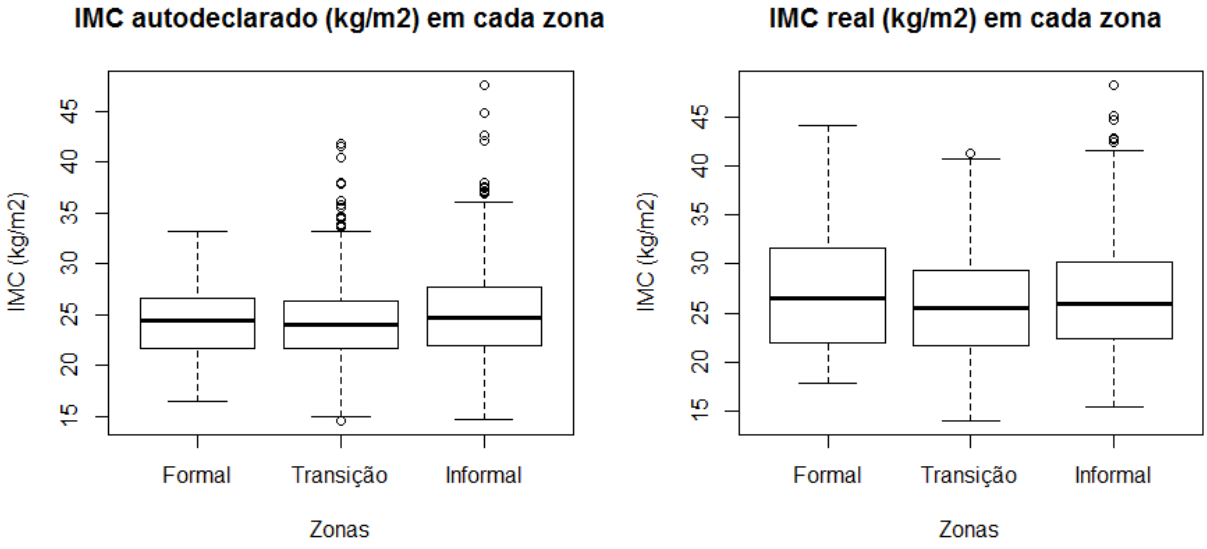


Figura B.7 - Boxplot correspondente ao IMC autodeclarado e ao IMC real. Representação do mínimo, primeiro quartil, mediana, terceiro quartil e máximo das variáveis quantitativas correspondentes ao IMC (kg/m^2) autodeclarado pelo inquirido e ao IMC real deste, por zona.

Anexo C: Código do R para a análise de mediação causal

```
################################
## Instalar o pacote mediation
################################

install.packages('mediation')
library("mediation")

################################
##Base de dados
################################

bd = read.table('bds.csv', header=TRUE, sep=';') #leitura da base de dados
str(bd) # variáveis com respectivos valores
summary(bd) #medidas sumárias das variáveis da base de dados
names(bd) #nomes das variáveis

################################
##base de dados mais pequena
################################

bd_small<-subset(bd,select=c(3:5,9:20,23:62,79:83))
dim(bd_small) #dimensão da amostra e número de variáveis
names(bd_small) #variáveis que constituem a bd_small
summary(bd_small) #medidas sumárias das variáveis da bd_small
casos_completos<-bd_small[complete.cases(bd_small),] #considerar apenas os casos completos
dim(casos_completos) #dimensão da amostra e número de variáveis
summary(casos_completos) #medidas sumárias dos casos_completos
names(bd_small) #variáveis que constituem os casos_completos
##################################################
##Estudo de Caso 1 - mediador: Habilitações literárias
##################################################

##criar variável binária referente a HL_Cat
D<-casos_completos$HL_Cat=='4' #com ensino secundário e curso médio
E<-casos_completos$HL_Cat=='5' #com curso superior
D<-as.numeric(D) #passagem a 1 e 0
E<-as.numeric(E) #passagem a 1 e 0
casos_completos$HL_BIN<-D+E # 1 - D+E -acima do ensino básico e inseriu-se em casos_completos, HL_BIN
names(casos_completos)

################################## Modelo sem covariáveis ########################################
###Para o total:
#modelo mediador:
modelo_mediador_1<-glm(HL_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"))
#modelo resposta:
modelo_resposta_1<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN,data=casos_completos,
                  family = binomial("probit"))
set.seed(2014)
mediacao_boot_1<- mediate(modelo_mediador_1, modelo_resposta_1, boot = TRUE, treat = "formal_informal",
                  mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_1) #Output com todos os efeitos

###Conforme o sexo:
```

```
#feminino
modelo_mediador_1fem<-glm(HL_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='1') #modelo mediador
modelo_resposta_1fem<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='1') #modelo resposta
set.seed(2014)
mediacao_boot_1F<- mediate(modelo_mediador_1fem, modelo_resposta_1fem, boot = TRUE,
                       treat = "formal_informal", mediator = "HL_BIN") #bootstrap não paramétrico

summary(mediacao_boot_1F) #Output com todos os efeitos

#masculino
modelo_mediador_1masc<-glm(HL_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='0') #modelo mediador
modelo_resposta_1masc<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='0') #modelo resposta
set.seed(2014)
mediacao_boot_1M<- mediate(modelo_mediador_1masc, modelo_resposta_1masc, boot = TRUE, treat = "formal_informal",
                   mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_1M) #Output com todos os efeitos


######################################### Modelo com  a covariável idade ######################################


###Para o total:
modelo_mediador_1id<-glm(HL_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit")) #modelo
                                                                                                           #mediador
modelo_resposta_1id<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade,data=casos_completos,
                     family = binomial("probit")) #modelo da resposta
set.seed(2014)
mediacao_boot_1I<- mediate(modelo_mediador_1id, modelo_resposta_1id, boot = TRUE, treat = "formal_informal",
                   mediator = "HL_BIN")#bootstrap não paramétrico
summary(mediacao_boot_1I) #Output com todos os efeitos

###Conforme o sexo:

#feminino

modelo_mediador_1femId<-glm(HL_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='1') #modelo mediador
modelo_resposta_1femId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='1') #modelo resposta
set.seed(2014)
mediacao_boot_1FI<- mediate(modelo_mediador_1femId, modelo_resposta_1femId, boot = TRUE,
                    treat = "formal_informal", mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_1FI) #Output com todos os efeitos


modelo_mediador_1id<-glm(HL_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit")) #modelo
                                                                                                           #mediador
modelo_resposta_1id<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade,data=casos_completos,
                     family = binomial("probit")) #modelo da resposta
set.seed(2014)
mediacao_boot_1I<- mediate(modelo_mediador_1id, modelo_resposta_1id, boot = TRUE, treat = "formal_informal",
                   mediator = "HL_BIN")#bootstrap não paramétrico
summary(mediacao_boot_1I) #Output com todos os efeitos

###Conforme o sexo:

#feminino

modelo_mediador_1femId<-glm(HL_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='1') #modelo mediador
modelo_resposta_1femId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='1') #modelo resposta
set.seed(2014)
mediacao_boot_1FI<- mediate(modelo_mediador_1femId, modelo_resposta_1femId, boot = TRUE,
                    treat = "formal_informal", mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_1FI) #Output com todos os efeitos
```

```
#masculino
modelo_mediador_1mascId<-glm(HL_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"),
                         subset = casos_completos$p1_sexo=='0') #modelo do mediador
modelo_resposta_1mascId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade,data=casos_completos,
                         family = binomial("probit"),subset = casos_completos$p1_sexo=='0') #modelo da resposta
set.seed(2014)
mediacao_boot_1MI<- mediate(modelo_mediador_1mascId, modelo_resposta_1mascId, boot = TRUE,
                    treat = "formal_informal",mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_1MI) #Output com todos os efeitos



######################### Modelo com  a covariável idade, sexo e tempo de residencia na zona ###################


###Para o total:

modelo_mediador_1_3cov<-glm(HL_BIN ~ formal_informal+idade+p1_sexo+p13_tempobairroAnos,data=casos_completos,
                        family = binomial("probit")) #modelo do mediador
modelo_resposta_1_3cov<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+HL_BIN+idade+p1_sexo+p13_tempobairroAnos,
                        data=casos_completos,family = binomial("probit")) #modelo da resposta
set.seed(2014)
mediacao_boot_13c<- mediate(modelo_mediador_1_3cov, modelo_resposta_1_3cov, boot = TRUE, treat = "formal_informal",
                    mediator = "HL_BIN") #bootstrap não paramétrico
summary(mediacao_boot_13c)#Output com todos os efeitos
################################################ Gráficos ############################################################

###Para o total de individuos:
par(mfrow=c(1,3))
plot(mediacao_boot_1, main = "Modelo sem covariáveis")
plot(mediacao_boot_1I, main = "Modelo com a covariável idade")
plot(mediacao_boot_13c, main = "Modelo com as covariáveis idade,
     sexo e tempo de residência na zona")

###Conforme o sexo:

par(mfrow=c(2,2))
#feminino
plot(mediacao_boot_1F, main = "Modelo sem covariáveis", sub="Sexo Feminino")
plot(mediacao_boot_1FI, main = "Modelo com a covariável idade", sub="Sexo Feminino")
#masculino
par(mfrow=c(1,2))
plot(mediacao_boot_1M, main = "Modelo sem covariáveis", sub="Sexo Masculino")
plot(mediacao_boot_1MI, main = "Modelo com a covariável idade", sub="Sexo Masculino")


#####################################################
##Estudo de Caso 2 - mediador: Situação Profissional
#####################################################

#criar variável binária referente a SitProf_Cat
A<-casos_completos$SitProf_Cat=='1' #trabalhadores
A<-as.numeric(A) #com 1 e 0
casos_completos$SitProf_Cat_BIN<-A # 1 - A (trabalhadores) e 0 - restantes(não trabalhadores) e inseriu-se
                                   #na base casos_completos
names(casos_completos)

################################### Modelo sem covariáveis ###########################################################

###Para o total:

modelo_mediador_2<-glm(SitProf_Cat_BIN~ formal_informal,data=casos_completos,family = binomial("probit"))
modelo_resposta_2<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN,data=casos_completos,
                   family = binomial("probit"))
set.seed(2014)
mediacao_boot_2<- mediate(modelo_mediador_2, modelo_resposta_2, boot = TRUE, treat = "formal_informal",
                  mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2)

###Conforme o sexo:
```

```
###Conforme o sexo:

#feminino
modelo_mediador_2fem<-glm(SitProf_Cat_BIN~ formal_informal,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='1')
modelo_resposta_2fem<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='1')
set.seed(2014)
mediacao_boot_2F<- mediate(modelo_mediador_2fem, modelo_resposta_2fem, boot = TRUE, treat = "formal_informal",
                   mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2F)


#masculino
modelo_mediador_2masc<-glm(SitProf_Cat_BIN~ formal_informal,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='0')
modelo_resposta_2masc<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN,data=casos_completos,
                       family = binomial("probit"),subset = casos_completos$p1_sexo=='0')
set.seed(2014)
mediacao_boot_2M<- mediate(modelo_mediador_2masc, modelo_resposta_2masc, boot = TRUE, treat = "formal_informal",
                    mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2M)


##################################### Modelo com  a covariável idade ##########################################
###Para o total:

modelo_mediador_2Id<-glm(SitProf_Cat_BIN~ formal_informal+idade,data=casos_completos,family = binomial("probit"))
modelo_resposta_2Id<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN+idade,data=casos_completos,
                     family = binomial("probit"))
set.seed(2014)
mediacao_boot_2I<- mediate(modelo_mediador_2Id, modelo_resposta_2Id, boot = TRUE, treat = "formal_informal",
                    mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2I)


###Conforme o sexo:

#feminino
modelo_mediador_2femId<-glm(SitProf_Cat_BIN~ formal_informal+idade,data=casos_completos,
                        family = binomial("probit"),subset = casos_completos$p1_sexo=='1')
modelo_resposta_2femId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN+idade,
                        data=casos_completos,family = binomial("probit"),subset = casos_completos$p1_sexo=='1')
set.seed(2014)
mediacao_boot_2FI<- mediate(modelo_mediador_2femId, modelo_resposta_2femId, boot = TRUE, treat = "formal_informal",
                    mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2FI)


#masculino
modelo_mediador_2mascId<-glm(SitProf_Cat_BIN~ formal_informal+idade,data=casos_completos,
                         family = binomial("probit"),subset = casos_completos$p1_sexo=='0')
modelo_resposta_2mascId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN+idade,
                         data=casos_completos,family = binomial("probit"),subset = casos_completos$p1_sexo=='0')
set.seed(2014)
mediacao_boot_2MI<- mediate(modelo_mediador_2mascId, modelo_resposta_2mascId, boot = TRUE,
                    treat = "formal_informal", mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_2MI)


########## Modelo com  a covariável idade, sexo e tempo de residencia na zona ##########

###Para o total:

modelo_mediador_2_3cov<-glm(SitProf_Cat_BIN~ formal_informal+idade+p1_sexo+p13_tempobairroAnos,
                        data=casos_completos,family = binomial("probit"))
modelo_resposta_2_3cov<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+SitProf_Cat_BIN+idade+
                        p1_sexo+p13_tempobairroAnos,data=casos_completos,family = binomial("probit"))
set.seed(2014)
mediacao_boot_23c<- mediate(modelo_mediador_2_3cov, modelo_resposta_2_3cov, boot = TRUE, treat = "formal_informal",
                    mediator = "SitProf_Cat_BIN")
summary(mediacao_boot_23c)

############################################## Gráficos ###########################################################
```

```
#Para o total:
par(mfrow=c(1,3))
plot(mediacao_boot_2, main = "Modelo sem covariáveis")
plot(mediacao_boot_2I, main = "Modelo com a covariável idade")
plot(mediacao_boot_23c, main = "Modelo com as covariáveis idade,
     sexo e tempo de residência na zona")

#Conforme o sexo:
par(mfrow=c(2,2))
#feminino
plot(mediacao_boot_2F, main = "Modelo sem covariáveis", sub="Sexo Feminino")
plot(mediacao_boot_2FI, main = "Modelo com a covariável idade", sub="Sexo Feminino")
#masculino
#par(mfrow=c(1,2))
plot(mediacao_boot_2M, main = "Modelo sem covariáveis", sub="Sexo Masculino")
plot(mediacao_boot_2MI, main = "Modelo com a covariável idade", sub="Sexo Masculino")



#####################################################
###Estudo de caso 3 - mediador: número de filhos
#####################################################

#criar variável binária referente a p11.1_Nfilhos
A<-casos_completos$p11.1_Nfilhos>='3' # Com 3 ou mais filhos
A<-as.numeric(A) #com 1 ou 0
casos_completos$Filhos_BIN<-A # 1 - >=3filhos e 0 <3 filhos e inseriu-se na base casos_completos
names(casos_completos)


################################### Modelo sem covariáveis ##################################################

###Para o total:

modelo_mediador_3<-glm(Filhos_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"))
#summary(modelo_mediador_9)
modelo_resposta_3<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN,data=casos_completos,
                  family = binomial("probit"))
set.seed(2014)
mediacao_boot_3<- mediate(modelo_mediador_3, modelo_resposta_3, boot = TRUE, treat = "formal_informal",
                  mediator = "Filhos_BIN")
summary(mediacao_boot_3)

###Conforme o sexo:

#feminino
modelo_mediador_3fem<-glm(Filhos_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"),
                      subset = casos_completos$p1_sexo=='1')
modelo_resposta_3fem<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='1')
set.seed(2014)
mediacao_boot_3F<- mediate(modelo_mediador_3fem, modelo_resposta_3fem, boot = TRUE, treat = "formal_informal",
                   mediator = "Filhos_BIN")
summary(mediacao_boot_3F)


#masculino
modelo_mediador_3masc<-glm(Filhos_BIN ~ formal_informal,data=casos_completos,family = binomial("probit"),
                       subset = casos_completos$p1_sexo=='0')
modelo_resposta_3masc<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN,data=casos_completos,
                      family = binomial("probit"),subset = casos_completos$p1_sexo=='0')
set.seed(2014)
mediacao_boot_3M<- mediate(modelo_mediador_3masc, modelo_resposta_3masc, boot = TRUE, treat = "formal_informal",
                   mediator = "Filhos_BIN")
summary(mediacao_boot_3M)


################################### Modelo com  a covariável idade ###########################################


###Para o total:
```

```
###Para o total:

modelo_mediador_3Id<-glm(Filhos_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"))
modelo_resposta_3Id<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN+idade,data=casos_completos,
                     family = binomial("probit"))
set.seed(2014)
mediacao_boot_3I<- mediate(modelo_mediador_3Id, modelo_resposta_3Id, boot = TRUE, treat = "formal_informal",
                   mediator = "Filhos_BIN")
summary(mediacao_boot_3I)



### Conforme o sexo:

#feminino
modelo_mediador_3femId<-glm(Filhos_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"),
                        subset = casos_completos$p1_sexo=='1')
modelo_resposta_3femId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN+idade,data=casos_completos,
                        family = binomial("probit"),subset = casos_completos$p1_sexo=='1')

set.seed(2014)
mediacao_boot_3FI<- mediate(modelo_mediador_3femId, modelo_resposta_3femId, boot = TRUE, treat = "formal_informal",
                    mediator = "Filhos_BIN")
summary(mediacao_boot_3FI)

#masculino
modelo_mediador_3mascId<-glm(Filhos_BIN ~ formal_informal+idade,data=casos_completos,family = binomial("probit"),
                         subset = casos_completos$p1_sexo=='0')
modelo_resposta_3mascId<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN+idade,data=casos_completos,
                         family = binomial("probit"),subset = casos_completos$p1_sexo=='0')

set.seed(2014)
mediacao_boot_3MI<- mediate(modelo_mediador_3mascId, modelo_resposta_3mascId, boot = TRUE,
                    treat = "formal_informal", mediator = "Filhos_BIN")
summary(mediacao_boot_3MI)



########## Modelo com  a covariável idade, sexo e tempo de residencia na zona ##########


###Para o total:

modelo_mediador_3_3cov<-glm(Filhos_BIN ~ formal_informal+idade+p1_sexo+p13_tempobairroAnos,data=casos_completos,
                        family = binomial("probit"))
modelo_resposta_3_3cov<-glm(ActivFisicaTempoLivre_Binaria~formal_informal+Filhos_BIN+idade+p1_sexo+
                        p13_tempobairroAnos,data=casos_completos,family = binomial("probit"))
set.seed(2014)
mediacao_boot_33c<- mediate(modelo_mediador_3_3cov, modelo_resposta_3_3cov, boot = TRUE, treat = "formal_informal",
                    mediator = "Filhos_BIN")
summary(mediacao_boot_33c)

######################################## Gráficos ############################################################

###Para o total:
par(mfrow=c(1,3))
plot(mediacao_boot_3, main = "Modelo sem covariáveis")
plot(mediacao_boot_3I, main = "Modelo com a covariável idade")
plot(mediacao_boot_33c, main = "Modelo com as covariáveis idade,
     sexo e tempo de residência na zona")

###Conforme o sexo:
par(mfrow=c(2,2))
#feminino
plot(mediacao_boot_3F, main = "Modelo sem covariáveis", sub="Sexo Feminino")
plot(mediacao_boot_3FI, main = "Modelo com a covariável idade", sub="Sexo Feminino")
#masculino
#par(mfrow=c(1,2))
plot(mediacao_boot_3M, main = "Modelo sem covariáveis", sub="Sexo Masculino")
plot(mediacao_boot_3MI, main = "Modelo com a covariável idade", sub="Sexo Masculino")
```

Anexo D: Código do R para a análise exploratória de dados

```
##########################################ANÁLISE EXPLORATÓRIA DOS DADOS##########################################

bd = read.table('bds.csv', header=TRUE, sep=';') #leitura da base de dados

#######################################
##Análise das variáveis quantitativas
#######################################


#################### Zona de aplicação do inquérito ###################################
indiv_zona<-table(bd$zona)
total<-indiv_zona[[1]]+indiv_zona[[2]]+indiv_zona[[3]]
tab.cont_1<-matrix(c(indiv_zona,total), ncol=4, nrow=1)
rownames(tab.cont_1)<-"Número de individuos"
colnames(tab.cont_1)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_1


###################### Divisão em Formal e Informal ########################################


#table(bd$formal_informal,bd$zona)
form_inform<-table(bd$formal_informal)
grupo_form<-table(bd$formal_informal[bd$zona=='1'])
grupo_trans<-table(bd$formal_informal[bd$zona=='2'])
grupo_inf<-table(bd$formal_informal[bd$zona=='3'])
total_form<-grupo_form[[1]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]
total<-total_form+total_trans+total_inf
tab.cont_2<-matrix(c(grupo_form,0,total_form, grupo_trans,total_trans,0,
            grupo_inf,total_inf,form_inform[[1]],form_inform[[2]],total),byrow=FALSE,ncol=4,nrow=3)
colnames(tab.cont_2)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
rownames(tab.cont_2)<-c("Zona Formal", "Zona Informal","Total")
tab.cont_2


################################## Sexo ###################################


#table(bd$p1_sexo,bd$zona)
sexo<-table(bd$p1_sexo)
grupo_form<-table(bd$p1_sexo[bd$zona=='1'])
grupo_trans<-table(bd$p1_sexo[bd$zona=='2'])
grupo_inf<-table(bd$p1_sexo[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_3<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,
             total_inf,sexo[[1]],sexo[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_3)<-c("Masculino","Feminino","Total")
colnames(tab.cont_3)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_3


######################### Idade recodificada em cinco categorias ##############################
```

```
#table(bd$idade_5cate,bd$zona)
idade_5cat<-table(bd$idade_5cate)
grupo_form<-table(bd$idade_5cate[bd$zona=='1'])
grupo_trans<-table(bd$idade_5cate[bd$zona=='2'])
grupo_inf<-table(bd$idade_5cate[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]+grupo_form[[5]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]+grupo_trans[[5]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]+grupo_inf[[5]]
total<-total_form+total_trans+total_inf
tab.cont_4<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,idade_5cat[[1]],
           idade_5cat[[2]],idade_5cat[[3]],idade_5cat[[4]],idade_5cat[[5]],total),byrow=FALSE,ncol=4,nrow=6)
rownames(tab.cont_4)<-c("<=29 anos","30-39 anos","40-49 anos","50-59 anos",">=60 anos","Total")
colnames(tab.cont_4)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_4


######################## Habilitações literárias categorizadas ##############################

#table(bd$HL_Cat,bd$zona)
hl_zona<-table(bd$HL_Cat)
grupo_form<-table(bd$HL_Cat[bd$zona=='1'])
grupo_trans<-table(bd$HL_Cat[bd$zona=='2'])
grupo_inf<-table(bd$HL_Cat[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]+grupo_form[[5]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]+grupo_trans[[5]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]+grupo_inf[[5]]
total<-total_form+total_trans+total_inf
tab.cont_5<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,hl_zona[[1]],
hl_zona[[2]],hl_zona[[3]],hl_zona[[4]],hl_zona[[5]],total),byrow=FALSE,ncol=4,nrow=6)
rownames(tab.cont_5)<-c("Sem escolaridade/Nunca", "Pré-escolar","Básico", "Secundário e Curso médio",
"Curso superior","Total")
colnames(tab.cont_5)<-c("Zona Formal", "Zona de Transição","Zona Informal","Total")
tab.cont_5


########################### Situação Profissional por zona ########################################

#table(bd$SitProf_Cat,bd$zona)
sitprof_zona<-table(bd$SitProf_Cat) #totais de cada categoria
grupo_form<-table(bd$SitProf_Cat[bd$zona=='1'])
grupo_trans<-table(bd$SitProf_Cat[bd$zona=='2'])
grupo_inf<-table(bd$SitProf_Cat[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]+grupo_form[[5]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]+grupo_trans[[5]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]+grupo_inf[[5]]
total<-total_form+total_trans+total_inf
tab.cont_6<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,sitprof_zona[[1]],
sitprof_zona[[2]],sitprof_zona[[3]],sitprof_zona[[4]],sitprof_zona[[5]],total),byrow=FALSE,ncol=4,nrow=6)
rownames(tab.cont_6)<-c("Trabalhador","Desempregado","Estudante","Reformado", "Doméstica","Total")
colnames(tab.cont_6)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_6


############################## Categorização do Estado civil ###############################################


#table(bd$EstadoCivil_Cat,bd$zona)
totalestado_zona<-table(bd$EstadoCivil_Cat) #apresenta o total de solteiros e nao solteiros
grupo_form<-table(bd$EstadoCivil_Cat[bd$zona=='1'])
grupo_trans<-table(bd$EstadoCivil_Cat[bd$zona=='2'])
grupo_inf<-table(bd$EstadoCivil_Cat[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_7<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,totalestado_zona[[1]],
totalestado_zona[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_7)<-c("Solteiro/Divorciado/Viúvo","Casado/União","Total")
colnames(tab.cont_7)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_7

############################## Se tem filhos ###############################################
```

```
filhos<-table(bd$p11_filhos) #totais de cada categoria (sim e não)
grupo_sfilhos<-table(bd$zona[bd$p11_filhos=='0'])
grupo_cfilhos<-table(bd$zona[bd$p11_filhos=='1'])
grupo_form<-table(bd$p11_filhos[bd$zona=='1'])
grupo_trans<-table(bd$p11_filhos[bd$zona=='2'])
grupo_inf<-table(bd$p11_filhos[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_8<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,filhos[[1]],
filhos[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_8)<-c("Não","Sim","Total")
colnames(tab.cont_8)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_8


################################## Há quanto tempo mora na zona #####################################


#table(bd$zona,bd$p13_tempobairroDN)
tempoZona_bin<-table(bd$p13_tempobairroDN)
grupo_form<-table(bd$p13_tempobairroDN[bd$zona=='1'])
grupo_trans<-table(bd$p13_tempobairroDN[bd$zona=='2'])
grupo_inf<-table(bd$p13_tempobairroDN[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_9<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
tempoZona_bin[[1]],tempoZona_bin[[2]],tempoZona_bin[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_9)<-c("Não desde o nascimento","Desde o nascimento", "Ns/Nr", "Total")
colnames(tab.cont_9)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_9


####################### Há quanto tempo mora na cidade da Praia ###################################

#table(bd$zona,bd$p14_tempoPraiaDN)
tempoPraia_bin<-table(bd$p14_tempoPraiaDN)
grupo_form<-table(bd$p14_tempoPraiaDN[bd$zona=='1'])
grupo_trans<-table(bd$p14_tempoPraiaDN[bd$zona=='2'])
grupo_inf<-table(bd$p14_tempoPraiaDN[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_10<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,tempoPraia_bin[[1]],
tempoPraia_bin[[2]],tempoPraia_bin[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_10)<-c("Não desde o nascimento","Desde o nascimento", "Ns/Nr","Total")
colnames(tab.cont_10)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_10


################################## Se gosta de viver na zona ########################################################
```

```
#table(bd$p17_gostaBairro,bd$zona)
gostabairro<-table(bd$p17_gostaBairro) #total de cada uma das 5 categorias
grupo_form<-table(bd$p17_gostaBairro[bd$zona=='1'])
grupo_trans<-table(bd$p17_gostaBairro[bd$zona=='2'])
grupo_inf<-table(bd$p17_gostaBairro[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]+grupo_form[[5]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]+grupo_trans[[5]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]+grupo_inf[[5]]
total<-total_form+total_trans+total_inf
tab.cont_11<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,
            total_inf,gostabairro[[1]],gostabairro[[2]],gostabairro[[3]],gostabairro[[4]],
            gostabairro[[5]],total),byrow=FALSE,ncol=4,nrow=6)
rownames(tab.cont_11)<-c("Nada","Pouco", "Indiferente", "Bastante", "Muito","Total")
colnames(tab.cont_11)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_11
```

```
######################### Tranquilidade ################################

#table(bd$p18_razaoBairroTranquil,bd$zona)
tranquilidade<-table(bd$p18_razaoBairroTranquil) #total de cada categoria (sim e não)
grupo_form<-table(bd$p18_razaoBairroTranquil[bd$zona=='1'])
grupo_trans<-table(bd$p18_razaoBairroTranquil[bd$zona=='2'])
grupo_inf<-table(bd$p18_razaoBairroTranquil[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_12<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            tranquilidade[[1]],tranquilidade[[2]],tranquilidade[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_12)<-c("Não","Sim","Ns/Nr", "Total")
colnames(tab.cont_12)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_12


################################# Motivos económicos ###################################

#table(bd$p18_razaoBairroEconom,bd$zona)
motivos_ec<-table(bd$p18_razaoBairroEconom) #total de cada categoria (sim e não)
grupo_form<-table(bd$p18_razaoBairroEconom[bd$zona=='1'])
grupo_trans<-table(bd$p18_razaoBairroEconom[bd$zona=='2'])
grupo_inf<-table(bd$p18_razaoBairroEconom[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_13<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            motivos_ec[[1]],motivos_ec[[2]],motivos_ec[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_13)<-c("Não","Sim", "Ns/Nr","Total")
colnames(tab.cont_13)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_13


######################################### Emprego no local ######################################
```

```
#table(bd$p18_razaoBairroEmprego,bd$zona)
emprego<-table(bd$p18_razaoBairroEmprego) #total de cada categoria (sim e não)
grupo_form<-table(bd$p18_razaoBairroEmprego[bd$zona=='1'])
grupo_trans<-table(bd$p18_razaoBairroEmprego[bd$zona=='2'])
grupo_inf<-table(bd$p18_razaoBairroEmprego[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_14<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            emprego[[1]],emprego[[2]],emprego[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_14)<-c("Não","Sim","Ns/nr", "Total")
colnames(tab.cont_14)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_14


############################## Relações familiares e habitações familiares ###################################

#table(bd$p18_razaoBairroFamilia,bd$zona)
familia<-table(bd$p18_razaoBairroFamilia) #total de cada categoria (sim e não)
grupo_form<-table(bd$p18_razaoBairroFamilia[bd$zona=='1'])
grupo_trans<-table(bd$p18_razaoBairroFamilia[bd$zona=='2'])
grupo_inf<-table(bd$p18_razaoBairroFamilia[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_15<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,familia[[1]],
            familia[[2]],familia[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_15)<-c("Não","Sim","Ns/Nr", "Total")
colnames(tab.cont_15)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_15



############################ Situação da zona em relaçao ao abandono ########################################

#table(bd$p25_bairroAbandono,bd$zona)
abandono<-table(bd$p25_bairroAbandono)
grupo_form<-table(bd$p25_bairroAbandono[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroAbandono[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroAbandono[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_16<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            abandono[[1]],abandono[[2]],abandono[[3]],abandono[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_16)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_16)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_16


########################### Situação da zona em relaçao aos assaltos/violência ############################

#table(bd$p25_bairroAssaltos,bd$zona)
assaltos<-table(bd$p25_bairroAssaltos)
grupo_form<-table(bd$p25_bairroAssaltos[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroAssaltos[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroAssaltos[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_17<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,assaltos[[1]],
assaltos[[2]],assaltos[[3]],assaltos[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_17)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_17)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_17


########################### Situação da zona em relaçao ao Custo de vida ################################
```

```
#table(bd$p25_bairroCustoVida,bd$zona)
cv<-table(bd$p25_bairroCustoVida)
grupo_form<-table(bd$p25_bairroCustoVida[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroCustoVida[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroCustoVida[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_18<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            cv[[1]],cv[[2]],cv[[3]],cv[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_18)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_18)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_18


############################## Situação da zona em relaçao ao desemprego ###################################


#table(bd$p25_bairroDesemprego,bd$zona)
desemprego<-table(bd$p25_bairroDesemprego)
grupo_form<-table(bd$p25_bairroDesemprego[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroDesemprego[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroDesemprego[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_19<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            desemprego[[1]],desemprego[[2]],desemprego[[3]],desemprego[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_19)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_19)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_19

############################## Situação da zona em relaçao à droga ################################
#table(bd$p25_bairroDroga,bd$zona)
droga<-table(bd$p25_bairroDroga)
grupo_form<-table(bd$p25_bairroDroga[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroDroga[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroDroga[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_20<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,droga[[1]],
            droga[[2]],droga[[3]],droga[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_20)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_20)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_20


######################## Situação da zona em relaçao à pobreza/exclusão social ################################
#table(bd$p25_bairroPobreza,bd$zona)
pobreza<-table(bd$p25_bairroPobreza)
grupo_form<-table(bd$p25_bairroPobreza[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroPobreza[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroPobreza[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_21<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,pobreza[[1]],
            pobreza[[2]],pobreza[[3]],pobreza[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_21)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_21)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_21


################################## Situação da zona em relaçao à prostituição ############################
```

```
#table(bd$p25_bairroProstituicao,bd$zona)
prostituicao<-table(bd$p25_bairroProstituicao)
grupo_form<-table(bd$p25_bairroProstituicao[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroProstituicao[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroProstituicao[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_22<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             prostituicao[[1]],prostituicao[[2]],prostituicao[[3]],prostituicao[[4]],total),
             byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_22)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_22)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_22


###################### Situação da zona em relaçao ao trânsito/acessibilidades ################################

#table(bd$p25_bairroTransito,bd$zona)
transito<-table(bd$p25_bairroTransito)
grupo_form<-table(bd$p25_bairroTransito[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroTransito[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroTransito[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_23<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,transito[[1]],
              transito[[2]],transito[[3]],transito[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_23)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_23)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_23


################ Situação da zona em relaçao à falta de civismo ################################

#table(bd$p25_bairroFaltaCivismo,bd$zona)
faltaCivismo<-table(bd$p25_bairroFaltaCivismo)
grupo_form<-table(bd$p25_bairroFaltaCivismo[bd$zona=='1'])
grupo_trans<-table(bd$p25_bairroFaltaCivismo[bd$zona=='2'])
grupo_inf<-table(bd$p25_bairroFaltaCivismo[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_24<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,faltaCivismo[[1]],
              faltaCivismo[[2]],faltaCivismo[[3]],faltaCivismo[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_24)<-c("Grave","Pouco Grave", "Sem gravidade", "Ns/Nr","Total")
colnames(tab.cont_24)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_24


######################### Qualidade de vida nos ultimos 5 anos na zona ############################

#table(bd$p26_qualidade5Anos,bd$zona)
qualidade_5<-table(bd$p26_qualidade5Anos) #total de cada categoria (melhorou, piorou, manteve-se igual, ns/nr)
grupo_form<-table(bd$p26_qualidade5Anos[bd$zona=='1'])
grupo_trans<-table(bd$p26_qualidade5Anos[bd$zona=='2'])
grupo_inf<-table(bd$p26_qualidade5Anos[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+grupo_form[[4]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_25<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             qualidade_5[[1]],qualidade_5[[2]],qualidade_5[[3]],qualidade_5[[4]],total),
             byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_25)<-c("Melhorou","Piorou", "Manteve-se igual", "Ns/Nr","Total")
colnames(tab.cont_25)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_25


######################### De forma geral viver na zona é... ###################################
```

```
#table(bd$p27_bairroGeral,bd$zona)
geral<-table(bd$p27_bairroGeral) #total de cada categoria (melhorou, piorou, manteve-se igual, ns/nr)
grupo_form<-table(bd$p27_bairroGeral[bd$zona=='1'])
grupo_trans<-table(bd$p27_bairroGeral[bd$zona=='2'])
grupo_inf<-table(bd$p27_bairroGeral[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]+0
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]+grupo_trans[[4]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]+grupo_inf[[4]]
total<-total_form+total_trans+total_inf
tab.cont_26<-matrix(c(grupo_form,0,total_form, grupo_trans,total_trans,grupo_inf,total_inf,geral[[1]],
            geral[[2]],geral[[3]],geral[[4]],total),byrow=FALSE,ncol=4,nrow=5)
rownames(tab.cont_26)<-c("Bom","Satisfatório", "Mau", "Ns/Nr","Total")
colnames(tab.cont_26)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_26


############################## Mais espaços desportivos por zona ##################################

#table(bd$p28_necessarioDesportivos,bd$zona)
espDesportivos<-table(bd$p28_necessarioDesportivos) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioDesportivos[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioDesportivos[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioDesportivos[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_27<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             espDesportivos[[1]],espDesportivos[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_27)<-c("Não","Sim","Total")
colnames(tab.cont_27)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_27


#################### Mais jardins e espaços verdes por zona ########################################

#table(bd$p28_necessarioJardins,bd$zona)
jardins<-table(bd$p28_necessarioJardins) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioJardins[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioJardins[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioJardins[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_28<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,jardins[[1]],
            jardins[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_28)<-c("Não","Sim","Total")
colnames(tab.cont_28)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_28


################## Maior segurança por zona ###################################
```

```
#table(bd$p28_necessarioSeguranca,bd$zona)
seguranca<-table(bd$p28_necessarioSeguranca) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioSeguranca[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioSeguranca[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioSeguranca[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_29<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             seguranca[[1]],seguranca[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_29)<-c("Não","Sim","Total")
colnames(tab.cont_29)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_29


################## Mais espaços públicos por zona ############################################
#table(bd$p28_necessarioEspPublicos,bd$zona)
espPub<-table(bd$p28_necessarioEspPublicos) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioEspPublicos[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioEspPublicos[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioEspPublicos[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_30<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,espPub[[1]],
             espPub[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_30)<-c("Não","Sim","Total")
colnames(tab.cont_30)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_30


############## Mais limpeza espaços públicos por zona ########################################
#table(bd$p28_necessarioLimpeza,bd$zona)
limpeza<-table(bd$p28_necessarioLimpeza) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioLimpeza[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioLimpeza[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioLimpeza[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_31<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,limpeza[[1]],
             limpeza[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_31)<-c("Não","Sim","Total")
colnames(tab.cont_31)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_31


######################## Mais transportes públicos por zona ################################
```

```
#table(bd$p28_necessarioTransportes,bd$zona)
transportes<-table(bd$p28_necessarioTransportes) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioTransportes[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioTransportes[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioTransportes[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_32<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             transportes[[1]],transportes[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_32)<-c("Não","Sim", "Total")
colnames(tab.cont_32)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_32


##################### Melhores acessibilidades por zona ########################################

#table(bd$p28_necessarioAcessibilidade,bd$zona)
acess<-table(bd$p28_necessarioAcessibilidade) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioAcessibilidade[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioAcessibilidade[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioAcessibilidade[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_33<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             acess[[1]],acess[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_33)<-c("Não","Sim","Total")
colnames(tab.cont_33)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_33


####################### Melhor ambiente no geral por zona ###################################

#table(bd$p28_necessarioAmbienteGeral,bd$zona)
ambGeral<-table(bd$p28_necessarioAmbienteGeral) #total de cada categoria (sim e não)
grupo_form<-table(bd$p28_necessarioAmbienteGeral[bd$zona=='1'])
grupo_trans<-table(bd$p28_necessarioAmbienteGeral[bd$zona=='2'])
grupo_inf<-table(bd$p28_necessarioAmbienteGeral[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_34<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
            ambGeral[[1]],ambGeral[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_34)<-c("Não","Sim", "Total")
colnames(tab.cont_34)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_34


############################## Dieta prescrita por zona ################################################
```

```
#table(bd$p45_dieta,bd$zona)
dieta<-table(bd$p45_dieta) #total de cada categoria (sim e não)
grupo_form<-table(bd$p45_dieta[bd$zona=='1'])
grupo_trans<-table(bd$p45_dieta[bd$zona=='2'])
grupo_inf<-table(bd$p45_dieta[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_35<-matrix(c(grupo_form,total_form, grupo_trans,0, total_trans,grupo_inf,0,
             total_inf,dieta[[1]],dieta[[2]],dieta[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_35)<-c("Não","Sim", "Ns/Nr","Total")
colnames(tab.cont_35)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_35


########################### se faz atividade fisica no trabalho #######################################
#table(bd$ActivFisicaTrabalho_Binaria,bd$zona)
actfisica_bin<-table(bd$ActivFisicaTrabalho_Binaria) #total de cada categoria (sim e não)
grupo_form<-table(bd$ActivFisicaTrabalho_Binaria[bd$zona=='1'])
grupo_trans<-table(bd$ActivFisicaTrabalho_Binaria[bd$zona=='2'])
grupo_inf<-table(bd$ActivFisicaTrabalho_Binaria[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_36<-matrix(c(grupo_form,total_form, grupo_trans,total_trans,grupo_inf,total_inf,
             actfisica_bin[[1]],actfisica_bin[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_36)<-c("Não","Sim","Total")
colnames(tab.cont_36)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_36


########################### se faz ou nao atividade fisica no tempo livre por zona ########################

#table(bd$ActivFisicaTempoLivre_Binaria,bd$zona)
actfisicalivre_bin<-table(bd$ActivFisicaTempoLivre_Binaria) #total de cada categoria (sim e não)
grupo_form<-table(bd$ActivFisicaTempoLivre_Binaria[bd$zona=='1'])
grupo_trans<-table(bd$ActivFisicaTempoLivre_Binaria[bd$zona=='2'])
grupo_inf<-table(bd$ActivFisicaTempoLivre_Binaria[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_37<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             actfisicalivre_bin[[1]],actfisicalivre_bin[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_37)<-c("Não","Sim","Total")
colnames(tab.cont_37)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_37


######################## Se anda a pe ou bicicleta pelo menos 10 min por zona ################################
```

```
#table(bd$p78_PeBicicleta,bd$zona)
pe_bic<-table(bd$p78_PeBicicleta) #total de cada categoria (sim e não e ns/nr e 99)
grupo_form<-table(bd$p78_PeBicicleta[bd$zona=='1'])
grupo_trans<-table(bd$p78_PeBicicleta[bd$zona=='2'])
grupo_inf<-table(bd$p78_PeBicicleta[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_38<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,
             total_inf,pe_bic[[1]],pe_bic[[2]],pe_bic[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_38)<-c("Não","Sim", "Ns/Nr","Total")
colnames(tab.cont_38)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_38


######################### Se anda de carro como forma de deslocação por zona ################################

#table(bd$p79_Carro,bd$zona)
carro<-table(bd$p79_Carro) #total de cada categoria (sim, não, ns/nr)
grupo_form<-table(bd$p79_Carro[bd$zona=='1'])
grupo_trans<-table(bd$p79_Carro[bd$zona=='2'])
grupo_inf<-table(bd$p79_Carro[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_39<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,carro[[1]],
             carro[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_39)<-c("Não","Sim", "Total")
colnames(tab.cont_39)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_39


################# Se anda de transporte publico como forma de deslocação por zona ############################


#table(bd$p79_TransportePublico,bd$zona)
trans_pub<-table(bd$p79_TransportePublico) #total de cada categoria (sim e não)
grupo_form<-table(bd$p79_TransportePublico[bd$zona=='1'])
grupo_trans<-table(bd$p79_TransportePublico[bd$zona=='2'])
grupo_inf<-table(bd$p79_TransportePublico[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_40<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             trans_pub[[1]],trans_pub[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_40)<-c("Não","Sim","Total")
colnames(tab.cont_40)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_40


######################### Se anda de taxi como forma de deslocação por zona ################################
```

```
#table(bd$p79_Taxi,bd$zona)
taxi<-table(bd$p79_Taxi) #total de cada categoria (sim e não)
grupo_form<-table(bd$p79_Taxi[bd$zona=='1'])
grupo_trans<-table(bd$p79_Taxi[bd$zona=='2'])
grupo_inf<-table(bd$p79_Taxi[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_41<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
            taxi[[1]],taxi[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_41)<-c("Não","Sim","Total")
colnames(tab.cont_41)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_41


############################## Se frequenta ginasio por zona ######################################

#table(bd$p82_PraticaGinasio,bd$zona)
ginasio<-table(bd$p82_PraticaGinasio) #total de cada categoria (sim e não - nao tem ns/nr)
grupo_form<-table(bd$p82_PraticaGinasio[bd$zona=='1'])
grupo_trans<-table(bd$p82_PraticaGinasio[bd$zona=='2'])
grupo_inf<-table(bd$p82_PraticaGinasio[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_42<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             ginasio[[1]],ginasio[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_42)<-c("Não","Sim","Total")
colnames(tab.cont_42)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_42


###################### Se frequenta clube desportivo por zona ##################################
#table(bd$p82_PraticaClubeDesportivo,bd$zona)
clubedesp<-table(bd$p82_PraticaClubeDesportivo) #total de cada categoria (sim e não - nr/ns sem nada
grupo_form<-table(bd$p82_PraticaClubeDesportivo[bd$zona=='1'])
grupo_trans<-table(bd$p82_PraticaClubeDesportivo[bd$zona=='2'])
grupo_inf<-table(bd$p82_PraticaClubeDesportivo[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_43<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             clubedesp[[1]],clubedesp[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_43)<-c("Não","Sim","Total")
colnames(tab.cont_43)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_43


################### Se frequenta piscina por zona ###################################
```

```
#table(bd$p82_PraticaPiscina,bd$zona)
piscina<-table(bd$p82_PraticaPiscina) #total de cada categoria (sim e não - nr/ns sem nada)
grupo_form<-table(bd$p82_PraticaPiscina[bd$zona=='1'])
grupo_trans<-table(bd$p82_PraticaPiscina[bd$zona=='2'])
grupo_inf<-table(bd$p82_PraticaPiscina[bd$zona=='3'])
total_form<-grupo_form[[1]]+0
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_44<-matrix(c(grupo_form,0,total_form, grupo_trans, total_trans,grupo_inf,total_inf,piscina[[1]],
            piscina[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_44)<-c("Não","Sim","Total")
colnames(tab.cont_44)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_44


########################### Se frequenta espaços publicos por zona ########################################
#table(bd$p82_PraticaEspacoPublico,bd$zona)
espacospub<-table(bd$p82_PraticaEspacoPublico) #total de cada categoria (sim e não)
grupo_form<-table(bd$p82_PraticaEspacoPublico[bd$zona=='1'])
grupo_trans<-table(bd$p82_PraticaEspacoPublico[bd$zona=='2'])
grupo_inf<-table(bd$p82_PraticaEspacoPublico[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_45<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
            espacospub[[1]],espacospub[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_45)<-c("Não","Sim","Total")
colnames(tab.cont_45)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_45



######################### Se frequenta piscina ou outro por zona ########################################
#table(bd$p82_PraticaOutroEPiscina,bd$zona)
pisc_outro<-table(bd$p82_PraticaOutroEPiscina) #total de cada categoria (sim e não)
grupo_form<-table(bd$p82_PraticaOutroEPiscina[bd$zona=='1'])
grupo_trans<-table(bd$p82_PraticaOutroEPiscina[bd$zona=='2'])
grupo_inf<-table(bd$p82_PraticaOutroEPiscina[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_46<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             pisc_outro[[1]],pisc_outro[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_46)<-c("Não","Sim","Total")
colnames(tab.cont_46)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_46


############## classificaçao IMC autodeclarado em 3 categorias ###################################

#table(bd$IMCautocatis_Rec1,bd$zona)
IMCauto_catRec1<-table(bd$IMCautocatis_Rec1) #total de cada categoria
grupo_form<-table(bd$IMCautocatis_Rec1[bd$zona=='1'])
grupo_trans<-table(bd$IMCautocatis_Rec1[bd$zona=='2'])
grupo_inf<-table(bd$IMCautocatis_Rec1[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_47<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
            IMCauto_catRec1[[1]],IMCauto_catRec1[[2]],IMCauto_catRec1[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_47)<-c("Baixo Peso","Peso Normal", "Excesso de Peso e Obesidade","Total")
colnames(tab.cont_47)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_47


################ Se permite realizar medidas ###################################
```

```
#table(bd$p94_autor1,bd$zona)
medidas<-table(bd$p94_autor1) #total de cada categoria (sim e não)
grupo_form<-table(bd$p94_autor1[bd$zona=='1'])
grupo_trans<-table(bd$p94_autor1[bd$zona=='2'])
grupo_inf<-table(bd$p94_autor1[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]
total<-total_form+total_trans+total_inf
tab.cont_48<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
             medidas[[1]],medidas[[2]],total),byrow=FALSE,ncol=4,nrow=3)
rownames(tab.cont_48)<-c("Não","Sim","Total")
colnames(tab.cont_48)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_48


##################### Classificação do IMC real em 3 categorias ###################################

#table(bd$IMCrealcatis_Rec1,bd$zona)
IMCreal_catRec1<-table(bd$IMCrealcatis_Rec1) #total de cada categoria
grupo_form<-table(bd$IMCrealcatis_Rec1[bd$zona=='1'])
grupo_trans<-table(bd$IMCrealcatis_Rec1[bd$zona=='2'])
grupo_inf<-table(bd$IMCrealcatis_Rec1[bd$zona=='3'])
total_form<-grupo_form[[1]]+grupo_form[[2]]+grupo_form[[3]]
total_trans<-grupo_trans[[1]]+grupo_trans[[2]]+grupo_trans[[3]]
total_inf<-grupo_inf[[1]]+grupo_inf[[2]]+grupo_inf[[3]]
total<-total_form+total_trans+total_inf
tab.cont_49<-matrix(c(grupo_form,total_form, grupo_trans, total_trans,grupo_inf,total_inf,
            IMCreal_catRec1[[1]],IMCreal_catRec1[[2]],IMCreal_catRec1[[3]],total),byrow=FALSE,ncol=4,nrow=4)
rownames(tab.cont_49)<-c("Baixo Peso","Peso Normal", "Excesso de Peso e Obesidade","Total")
colnames(tab.cont_49)<-c("Zona Formal","Zona de Transição","Zona Informal","Total")
tab.cont_49
```

```
#######################################
##Análise das variáveis qualitativas
#######################################


################################## Tamanho do Agregado ########################################

#table(bd$Agregado)#total
#table(bd$Agregado[bd$zona==1]) #zona 1
#table(bd$Agregado[bd$zona==2]) #zona 2
#table(bd$Agregado[bd$zona==3]) #zona 3

#Estatísticas descritivas
summary(bd$Agregado)
summary(bd$Agregado[bd$zona==1]) #zona 1
summary(bd$Agregado[bd$zona==2]) #zona 2
summary(bd$Agregado[bd$zona==3]) #zona 3


################################## Número de adultos incluindo o próprio ##################################

#table(bd$p12.2_num_adulto) #total
#table(bd$p12.2_num_adulto[bd$zona==1]) #zona 1
#table(bd$p12.2_num_adulto[bd$zona==2]) #zona 2
#table(bd$p12.2_num_adulto[bd$zona==3]) #zona 3
#Estatísticas descritivas
summary(bd$p12.2_num_adultos) #total
summary(bd$p12.2_num_adultos[bd$zona==1]) #zona 1
summary(bd$p12.2_num_adultos[bd$zona==2]) #zona 2
summary(bd$p12.2_num_adultos[bd$zona==3]) #zona 3


#Boxplot conjunto conforme as zonas
par(mfrow=c(1,2))
boxplot(bd$Agregado ~ bd$zona, xlab="Zonas", main="Tamanho do agregado familiar,
        em cada zona",ylab="Nrº indivíduos",names=c("Formal","Transição","Informal"))
boxplot(bd$p12.2_num_adultos ~ bd$zona, xlab="Zonas", main="Quantidade de adultos, incluindo o próprio,
        em cada zona",ylab="Nrº adultos",names=c("Formal","Transição","Informal"))


####################### Quantidade de filhos ###################################

#table(bd$p11.1_Nfilhos) #total
#table(bd$p11.1_Nfilhos[ bd$zona==1]) #zona 1
#table(bd$p11.1_Nfilhos[bd$zona==2]) #zona 2
#table(bd$p11.1_Nfilhos[bd$zona==3]) #zona 3

#Estatisticas descritivas
summary(bd$p11.1_Nfilhos) #total
summary(bd$p11.1_Nfilhos[bd$zona==1]) #zona 1
summary(bd$p11.1_Nfilhos[bd$zona==2]) #zona 2
summary(bd$p11.1_Nfilhos[bd$zona==3]) #zona 3
```

```
#boxplot conforme as zonas
boxplot(bd$p11.1_Nfilhos ~ bd$zona, xlab="Zonas", main="Quantidade de filhos em cada zona",
        ylab="Nrº filhos",names=c("Formal","Transição","Informal"))

########################### Casa onde vive, dispõe de quantos quartos ##############################

#table(bd$p10_alojQ)
#table(bd$p10_alojQ[bd$zona==1]) #zona 1
#table(bd$p10_alojQ[bd$zona==2]) #zona 2
#table(bd$p10_alojQ[bd$zona==3]) #zona 3

#Estatísticas descritivas
summary(bd$p10_alojQ)
summary(bd$p10_alojQ[bd$zona==1]) #zona 1
summary(bd$p10_alojQ[bd$zona==2]) #zona 2
summary(bd$p10_alojQ[bd$zona==3]) #zona 3

#boxplot conforme as zonas
boxplot(bd$p10_alojQ ~ bd$zona, xlab="Zonas", main="Quantidade de quartos numa casa, em cada zona",
ylab="Nrº quartos",names=c("Formal","Transição","Informal"))

####################### Em média, quantas horas passa sentado por dia, por zona ######################

#Estatisticas descritivas
summary(bd$p84_horasSentado)
summary(bd$p84_horasSentado[bd$zona=='1'])
summary(bd$p84_horasSentado[bd$zona=='2'])
summary(bd$p84_horasSentado[bd$zona=='3'])
########################### Em média, quantas horas dorme por dia ##################################

#Estatisticas descritivas
summary(bd$P85_horasdorm)
summary(bd$P85_horasdorm[bd$zona=='1'])
summary(bd$P85_horasdorm[bd$zona=='2'])
summary(bd$P85_horasdorm[bd$zona=='3'])



#boxplot conjunto, conforme as zonas
par(mfrow=c(1,2))
boxplot(bd$p84_horasSentado ~ bd$zona, xlab="Zonas", main="Número médio de horas sentado por dia,
        em cada zona", ylab="Horas",names=c("Formal","Transição","Informal"))
boxplot(bd$P85_horasdorm ~ bd$zona, xlab="Zonas", main="Número médio de horas de sono por dia,
        em cada zona",ylab="Horas",names=c("Formal","Transição","Informal"))

############################### Altura real (m) por zona ######################################3

#table(bd$p95.1_alturareal)
#table(bd$p95.1_alturareal,bd$zona)

#Estatisticas descritivas
summary(bd$p95.1_alturareal) #total
summary(bd$p95.1_alturareal[bd$zona=='1']) #zona 1
summary(bd$p95.1_alturareal[bd$zona=='2']) #zona 2
summary(bd$p95.1_alturareal[bd$zona=='3']) #zona 3
```

```
#boxplot conforme as zonas
par(mfrow=c(1,1))
boxplot(bd$p95.1_alturareal ~ bd$zona, xlab="Zonas", main="Altura real (m) em cada zona",
        ylab="Altura (m)",names=c("Formal","Transição","Informal"))

################### Qual o peso habitual sem roupa nem sapatos ############################

#table(bd$p92_pesoauto,bd$zona)

#estatisticas descritivas
summary(bd$p92_pesoauto)
summary(bd$p92_pesoauto[bd$zona=='1'])
summary(bd$p92_pesoauto[bd$zona=='2'])
summary(bd$p92_pesoauto[bd$zona=='3'])

######################### Peso desejado ######################################

#table(bd$p93_pesodese)
#table(bd$p93_pesodese,bd$zona)

#estatisticas descritivas
summary(bd$p93_pesodese)  #total
summary(bd$p93_pesodese[bd$zona=='1']) #Zona 1
summary(bd$p93_pesodese[bd$zona=='2']) #Zona 2
summary(bd$p93_pesodese[bd$zona=='3']) #Zona 3

######################### Peso Real (Kg) ######################################

#table(bd$p95_pesoreal)
#table(bd$p95_pesoreal,bd$zona)

#Estatisticas descritivas
summary(bd$p95_pesoreal) #total
summary(bd$p95_pesoreal[bd$zona=='1']) #zona 1
summary(bd$p95_pesoreal[bd$zona=='2']) #zona 2
summary(bd$p95_pesoreal[bd$zona=='3']) #zona 3

#boxplot conjunto conforme as zonas
par(mfrow=c(1,3))
boxplot(bd$p92_pesoauto ~ bd$zona, xlab="Zonas", main="Peso autodeclarado (kg) em cada zona",
ylab="Peso (kg)",names=c("Formal","Transição","Informal"))
boxplot(bd$p93_pesodese ~ bd$zona, xlab="Zonas", main="Peso desejado (kg) em cada zona",
ylab="Peso (kg)",names=c("Formal","Transição","Informal"))
boxplot(bd$p95_pesoreal ~ bd$zona, xlab="Zonas", main="Peso real (kg) em cada zona",
ylab="Peso (kg)",names=c("Formal","Transição","Informal"))

############################## IMC autodeclarado ############################

#table(bd$IMC_auto,bd$zona)

#estatisticas descritivas
summary(bd$IMC_auto)
summary(bd$IMC_auto[bd$zona=='1'])
summary(bd$IMC_auto[bd$zona=='2'])
summary(bd$IMC_auto[bd$zona=='3'])

############################## IMC real (Kg/m2) por zona ######################
```