# UNIVERSIDADE ABERTA


# Modelos de Regressão Linear: "Fatores que Influenciam a Esperança de Vida em Angola"

**Francisco Tchissaquila Chimuco**

**Mestrado em Estatística, Matemática e Computação na área de Especialização de Estatística Computacional**

**2015**

UNIVERSIDADE ABERTA


# Modelos de Regressão Linear: "Fatores que Influenciam a Esperança de Vida em Angola"

**Francisco Tchissaquila Chimuco**

**Mestrado em Estatística, Matemática e Computação na área de Especialização de Estatística Computacional**

**Dissertação orientada pelo Professor Doutor Amilcar Manuel do Rosário Oliveira e pela Professora Doutora Célia Maria Pinto Nunes**

**2015**

# Resumo

A esperança de vida é o número médio de anos que um indivíduo viverá a partir do nascimento, considerando o nível e estrutura de mortalidade por idade observados naquela população. Para o cálculo da esperança de vida ao nascer leva-se em consideração não apenas os riscos de morte na primeira idade – mortalidade infantil -, mas todo o histórico de mortalidade de crianças, adolescentes, jovens, adultos e idosos. Sendo uma síntese da mortalidade ao longo de todo o ciclo de vida dos indivíduos, a esperança de vida é o indicador empregado para mensurar as dimensões humanas no índice de desenvolvimento, qual seja, direito a uma vida longa e saudável. Isso porque, em cada um dos grupos etários os indivíduos estão sujeitos a diferentes riscos de mortalidade, estabelecendo distintas causas principais de mortalidade. O objectivo do trabalho é identificar as principais causas que contribuem para a diminuição dos índices da esperança de vida em Angola, construir e propor um modelo de regressão linear múltipla para se prevenir fenómenos futuros que permitam tirar conclusões relativamente aos factores que contribuem para o aumento da esperança de vida em Angola. Para tal recorreu-se aos dados estatísticos publicados em relatórios pelos órgãos oficiais angolanos como o Instituto Nacional de Estatística e internacionais como a Organização das Nacões Unidas entre outros, a fim de se fazer a recolha dos dados que permitissem caracterizar o fenómeno em estudo em Angola. Por outro lado, fez-se o estudo da regressão linear simples e múltipla no sentido de se propor um modelo estatístico que justifique os dados. Calculou-se ainda a correlação entre as diversas variáveis no sentido de identificar as que possuem forte correlação positiva com a esperança de vida ao nascer em Angola. Com a utilização do *software IBM SPSS 20 for Windows* foram calculadas as estatísticas descritivas, os coeficientes de correlação de Pearson entre as variáveis independentes e a esperança de vida e as análises de regressão linear múltipla, que permitiram gerar modelos preditivos para cada uma das regiões angolanas.

**Palavras-chave:** Esperança de vida, Regressão Linear Múltipla, Saúde Colectiva, Equidade, Variáveis Sociodemográficas.

# Abstract

Life expectancy is the average of years a person will live from birth, considering the level and mortality structure observed in a particular population. To calculate life expectancy at birth one takes into account not only the risk of death in infancy - infant mortality - but the whole history of mortality of children, adolescents, youth, adults and elderly people. Being a synthesis of mortality throughout the life cycle of individuals, life expectancy is the indicator used to measure the human dimensions in development index, that is, the right to a long and healthy life. This is because in each of the age groups individuals are subject to different mortality risks, establishing distinct major causes of mortality. The objective of this study is to identify the main causes contributing to the reduction of rates of life expectancy in Angola, create and propose a multiple linear regression model to prevent future events that allow conclusions regarding the factors contributing to the increase life expectancy in Angola. To collect data we resort to the statistics published in reports by Angolan official bodies such as the National Institute of Statistics and international sources as the United Nations and others, making it possible to characterize the phenomenon under study in Angola. On the other hand, a study of simple and multiple linear regression was carried out in order to propose a statistical model to justify the data. It is further calculated the correlation between the different variables in order to identify those with strong positive correlation with the hope life expectancy at birth in Angola.. Applying IBM SPSS 20 for Windows descriptive statistics were calculated, the Pearson correlation coefficients between the independent variables, life expectancy and multiple linear regression analyzes which helped to create predictive models for each of the Angolan regions.



**Keywords**: Life expectancy, Multiple Linear Regression, Collective Health, Equity, Sociodemographic variables.

## Dedicatória

À minha esposa Paulina Chimuco e aos meus filhos Alex, Franklin, Jesuina e Adelaide.

## Agradecimentos

Em primeiro lugar agradeço a Deus todo poderoso pela força e coragem para não desistir.

Aos meus orientadores Professor Doutor Amílcar de Oliveira e a Professora Doutora Célia Nunes pelas incansáveis e sábias orientações no sentido de continuar e nunca desistir nessa árdua tarefa de investigação.

Aos meus pais e irmãos por nunca deixarem faltar amor, carinho e confiança.

À minha esposa, Paulina Chimuco pela paciência e prestatividade no desenvolvimento desta investigação.

A todos aqueles que direta ou indiretamente contribuíram para a realização desta investigação.

# Índice

# Índice de Tabelas

# Índice de Gráficos

# Índices de Figuras

## Lista de abreviaturas, siglas, acrónimos e símbolos matemáticos

**ANOVA** (Análise de Variância)

**CV** (Coeficiente de Variação)

**CIA** (Central Inteligency Agency)

**IBEP** (Inquérito Integrado Sobre o Bem-Estar da População

**INE** (Instituto Nacional de Estatística)

**MINSA** (Ministério da Saúde)

**N** (Distribuição Normal)

**NID** (Distribuição Normal e Independente)

**OMS** (Organização Mundial da Saúde)

**ONU** (Organização das Nações Unidas)

**PNUD** (Programa das Nações Unidas para o Desenvolvimento)

**QUIBB** (Inquérito de Indicadores Básicos de Bem-Estar)

**SIDA** (Síndrome da Imunodeficiência Adquirida)

**UNICEF** (Fundo das Nações Unidas para Infância)

**VIH** (Virus da Imunodeficiência Humana)

$\sqrt{QME}$ (Erro padrão dos resíduos)

**WHO** (World Health Organization)

$SS_R$ (Soma dos quadrados da regressão)

$SS_E$ (Soma dos quadrados dos erros)

$SS_T$ (Soma dos quadrados total)

# 1 Introdução

O presente trabalho tem como objetivo apresentar os resultados mais importantes sobre os modelos de regressão lineares começando com o modelo de regressão linear simples e prosseguindo depois para modelos mais complexos de regressão linear múltipla. Tendo em conta a importância da validação dos pressupostos impostos ao erro do modelo, pretende-se também abordar a análise de resíduos. Com vista a mostrar a grande aplicabilidade destes modelos pretende-se ainda apresentar algumas aplicações a dados reais recorrendo ao uso do *software* estatístico SPSS.

Para o efeito, serão apresentados os modelos teóricos e seus pressupostos e realizar-se-á inferência para os parâmetros dos modelos com a construção de intervalos de confiança e testes de hipóteses, nomeadamente o teste t-student e a análise de variância. Na parte da aplicação, e tendo em atenção a pesquisa estatística feita, utilizaremos os dados da Organização Mundial da Saúde e do Instituto Nacional de Estatística de Angola para propor um modelo de regressão linear simples ou múltipla.

## 1.1 Justificação

A correlação linear é um indicador indiscutível para a explicação de relações de causa – efeito e a análise de regressão linear forneceuma função matemática que descreve a relação a relação entre duas ou mais variáveis. Esta função/equação pode ser usada para estimar ou prever valores futuros e uma variável com base em volores conhecidos ou supostos, de uma ou mais variáveis relacionadas. Assim sendo, e tendo em conta que a esperança de vida num determinado país tem sido interpretada como resultado do efeito de vários fatores como a pobreza, o acesso aos serviços básicos de saúde, saneamento básico, educação, cultura e lazer, bem como os índices de violência, criminalidade, segurança, poluição do local onde vive a população, entre outros, não está no entanto claro quais destes fatores exercem maior influência neste fenómeno. Esta questão motivou-nos a levar em frente esta pesquisa no sentido de verificar quais as verdadeiras causas que estão por detrás deste fenómeno em Angola.

Segundo Hubert Chamone Gesser (2005, pag. 15) na sua tese de doutoramento, o conhecimento dos fatores que interferem na elevação ou na redução da esperança de vida mostram-se importantes, não somente pelas características da saúde pública, mas também pelas consequências socioeconómicas que afetam as populações. As políticas

públicas necessitam de um rigoroso planeamento para que os resultados a serem alcançados tragam o retorno esperado e desejado. O alcance desses resultados é dependente dos fundamentos científicos que são adotados na fase de planeamento do que deve ser feito, das intervenções a serem adotadas e das medidas a serem tomadas.

Do exposto acima e atendendo aos números divulgados pelas organizações internacionais no que concerne aos dados sociodemográficos relativos a Angola, fica claro que as autoridades angolanas não têm tido em conta tais resultados.

Pensamos serem estes motivos suficientes para analisar as causas que têm maior influência na esperança de vida dos países africanos no geral e em particular em Angola e propor um modelo de regressão linear que permita explicar e tentar prever fenómenos futuros.

## 1.2 Problema

Da análise feita aos relatórios da Organização Mundial de Saúde no Website oficial e dos dados disponíveis no Website da Agência Central de Inteligência Americana (CIA), constatamos que existe um fosso muito grande na esperança de vida entre os países ocidentais e/ou desenvolvidos e os países africanos no geral e em particular em Angola. Este facto levou-nos a questionar, o porquê desta disparidade e chegámos à conclusão que existe uma pergunta para a qual gostaríamos de encontrar uma resposta, ou seja qu**e** fatores condicionam com maior relevância a diminuição da esperança de vida ao nascer, em África e em particular em Angola e como contribuir para diminuir tal realidade.

## 1.3 Objectivos

### 1.3.1 Geral:

- Identificar as principais causas que contribuem para a diminuição dos índices da esperança de vida em Angola. Construir e propor um modelo de regressão linear para se prevenir fenómenos futuros que tirar conclusões relativamente aos fatores que concorrem para a diminuição da esperança de vida em Angola.

### 1.3.2 Específicos:

- Estudar o modelo de regressão linear simples e o modelo de regressão linear múltiplo;
- Realizar uma pesquisa documental a fim de identificar os fatores de risco relativamente à esperança de vida ao nascer e viver em Angola;

- Fazer o tratamento dos dados para verificar a correlação entre as diversas variáveis no sentido de identificar as que possuem forte correlação positiva com a esperança de vida ao nascer em Angola e propor um modelo de regressão linear simples ou múltipla no sentido de permitir prever e prevenir o fenómeno em estudo.

## 1.4 Hipótese

Pretendemos verificar, como referido anteriormente, quais os fatores que influenciam a esperança de vida em Angola. Nomeadamente, pretendemos averiguar se a taxa de mortalidade por VIH/SIDA, a taxa de mortalidade infantil (mortes/1.000 nascimentos normais), as camas hospitalares *per capita* (leitos /1.000 habitantes), o rendimento *per capita*, o tamanho da população, a taxa de alfabetização, o acesso a água canalizada e a inexistência/existência de apoio médico, condicionam aa esperança de vida ao nascer emem Angola (Gesser, 2005).

## 1.5 Resultados esperados

A esperança de vida de qualquer sociedade é um indicador muito importante e um bom indicador do Índice de Desenvolvimento Humano, condição necessária para se garantir os direitos e garantias fundamentais para as sociedades atuais. Nesta base, com esta investigação, pretendemos alcançar os seguintes resultados:

**Científico:** expectativa de apresentar argumentos e resultados acerca dos fatores que condicionam o aumento da esperança de vida ao nascer em Angola.

## 1.6 Estrutura do trabalho

Visando a apresentação do relatório de dissertação de maneira organizada e tendo em atenção a metodologia de investigação científica no que concerne a estruturação dos relatórios de dissertações, este relatório possuirá a seguinte estrutura para além da Introdução:

- Capítulo 2 - Fundamentação teórica, no qual se apresenta a teoria de base que sustenta a pesquisa, a revisão da literatura bem como a definição dos termos ou conceitos;
- Capítulo 3 – Regressão Linear: onde são apresentados os procedimentos estatísticos de correlação e de análise de regressão linear múltipla, que foram utilizados para a análise da base de dados;

- Capítulo 4 - Desenho metodológico:  no qual se abordam os itens, método de investigação, técnicas, delimitação do universo (descrição da população).
- Capítulo 5 – Aplicação da Análise de Regressão Linear: aqui apresentam-seas análises das correlações das variáveis e os modelos explicativos, resultantes da análise de regressão linear múltipla, da esperança de vida para cada uma das regiões angolanas;
- Capítulo 6 – Conclusões e sugestões, aqui apresentam-se as discussões do autor relativamente aos resultados encontrados no estudo e que têm maior impacto e/ou relevância científica no conhecimento da esperança de vida como solução do problema levantado através do estudo considerando a base de dados. Apresenta-se ainda neste capítulo uma abordagem com algumas indicações, de ordem prática, de intervenções na sociedade, de acordo com as conclusões do estudo. Ainda neste capítulo apresentaremos sugestões para trabalhos futuros.

# 2 Enquadramento teórico

## 2.1 A Esperança de Vida

A esperança de vida é definida, segundo o *site* oficial dos dados estatísticos de Portugal, como sendo o número médio de anos que uma pessoa pode esperar viver, mantendo-se as taxas de mortalidade por idades observadas no momento de referência.

Num artigo brasileiro a esperança de vida é definida como sendo o número médio de anos que um indivíduo viverá a partir do nascimento, considerando o nível e estrutura de mortalidade por idade observados naquela população. Ainda segundo a *wikipedia*, numa dada população, com base nos dados do Índice de Desenvolvimento Humano (IDH) de qualquer ano, a expectativa de vida ao nascer ou esperança de vida à nascença é o número médio de anos que um grupo de indivíduos nascidos no mesmo ano pode esperar viver, se mantidas, desde o seu nascimento, as taxas de mortalidade observadas no ano de observação. A esperança de vida no nascimento é também um indicador de qualidade de vida de um país, região ou localidade. Pode também ser utilizada para aferir o retorno de investimentos feitos na melhoria das condições de vida e para compor vários índices, tais como o IDH.

De acordo com os dados disponibilizados pela CIA (*Central Inteligency Agency*), os primeiros cinco países com maiores índices de esperança de vida em anos em 2013 são o Mónaco com 89.63, Macau com 84.63, Japão com 84.19, Singapura com 84.07 e San Marino com 83.12. O último lugar no *ranking* é ocupado pelo Chade com 49.97 anos em média e Angola ocupa o 203º lugar no Ranking, num total de 230 países, com 54.95 anos. É de realçar que houve alguma subida gradual em termos de esperança de vida em Angola de 15.95 anos em média, desde 1990 a 2013. Relativamente aos países africanos o país que se encontra melhor posicionado é a Argélia, com 76.18 anos em média, ocupando o 82º lugar no *ranking* de 2013 num total de 230.

Fazendo-se uma comparação com outros países em termos da esperança de vida ao nascer temos a salientar o facto de alguns países estarem próximos de Angola como são os casos do Mali, Botswana, Burkina Faso e Níger. Mas, não podemos deixar de realçar o facto de alguns países da mesma região e que supostamente se encontram em conflito e com menor PIB que Angola possuírem maior índice de esperança de vida, como são os casos da República Democrata do Congo com 56.14 e a República do Congo com 55.60.

Um indicador interessante nestes dados é o facto de as mulheres possuírem esperança de vida maior que os homens no geral e em particular em Angola. Em 2011 existiu uma diferença de 3 anos. Fazendo uma comparação, podemos notar que há uma tendência no sentido de haver algum equilíbrio nos próximos anos.

Vejamos a seguir a tabela abaixo que mostra a evolução da esperança de vida em Angola em 1990, 2000, 2011, 2012 e 2013, distribuído pelo sexo.

**Tabela 2.1** Evolução da Esperança de vida ao nascer em Angola

| | **Esperança de Vida ao Nascer (anos)** | | | |
|---|---|---|---|---|
| | | **Sexo** | | |
| **País** | **Ano** | **Masculino** | **Feminino** | **Média** |
| Angola | 1990 | 41 | 45 | 43 |
| | 2000 | 44 | 47 | 46 |
| | 2011 | 50 | 53 | 52 |
| | 2012 | 50 | 52 | 51 |
| | 2013 | 50 | 52 | 52 |

**Fonte:** OMS (2014)

A busca das causas das reduções ou incrementos na esperança de vida das pessoas, também denominado esperança de vida ao nascer, é uma das preocupações mais frequentes daqueles que estudam os relacionamentos entre os fenômenos biológicos e os fenômenos sociais (Gesser, 2005).

Thisted (2003), citado por Gesser (2005), confirma a existência da determinação social no processo saúde-doença, onde parece claro, através de relatos encontrados na literatura, que as condições de saúde das populações são fortemente afetadas por questões de natureza social, como o grau de escolaridade, o rendimento familiar, etc.

Ansari *et al*. (2003), citado por Gesser (2005), alertaram para o fato de muitos estudos em saúde ignorarem os fatores sociais como determinantes de saúde ou como fatores que se inter-relacionam com as variáveis biológicas. Os fatores sociais são muito relevantes para a elaboração de um modelo explicativo de uma doença, e não devem ser negligenciados quando se deseja a obtenção de intervenções em saúde pública baseadas em evidências.

Segundo Gesser (2005), os estudos que buscam os relacionamentos destes fenómenos procuram explicar, através de estudos epidemiológicos, com os mais variados desenhos (corte, prevalência, caso controlo, ecológico), o grau de impacto que esses determinantes sociais têm sobre a saúde das populações.

Um marco nos estudos sobre a determinação social nas condições de saúde das populações foi a teoria proposta por Hart, na Grã-Bretanha, em 1970. Esta teoria demonstra as razões pela qual as pessoas mais pobres têm menor acesso aos serviços de saúde (Hart, 1971, conforme refere Gesser, 2005). De acordo com Hart nas áreas com mais doentes e mortalidade elevada, há mais filas de espera, menor suporte hospitalar, obsolescência de equipamentos e carência de leitos e recursos humanos. Esses achados foram denominados "*the inverse care law*" ou lei dos cuidados inversos, relatando que a disponibilidade dos recursos de saúde estão em uma relação inversa se comparados com as reais necessidades das populações atendidas.

Victoria *et al.* (2000), citado por Gesser (2005), propuseram a "*inverse equity hypothesis*", ou hipótese da equidade inversa, para explicar que as intervenções e os programas de saúde inicialmente atingem as pessoas de nível socioeconómico mais alto para somente mais tarde, alcançarem as pessoas mais pobres. De acordo com os autores, somente após o alcance de níveis de morbidade e mortalidade baixos para os mais ricos é que os mais pobres passam a ganhar acesso às mesmas intervenções.

Para o conceito de equidade, adota-se o proposto por Mooney (1983) citado por Gesser (2005), a "igualdade de recursos para igual necessidade" tomando-se em consideração a estrutura da população, por sexos e grupos de idade, bem como outros fatores que influenciem as respectivas necessidades. Logo, ao se buscar a equidade faz-se a promoção de justiça social.

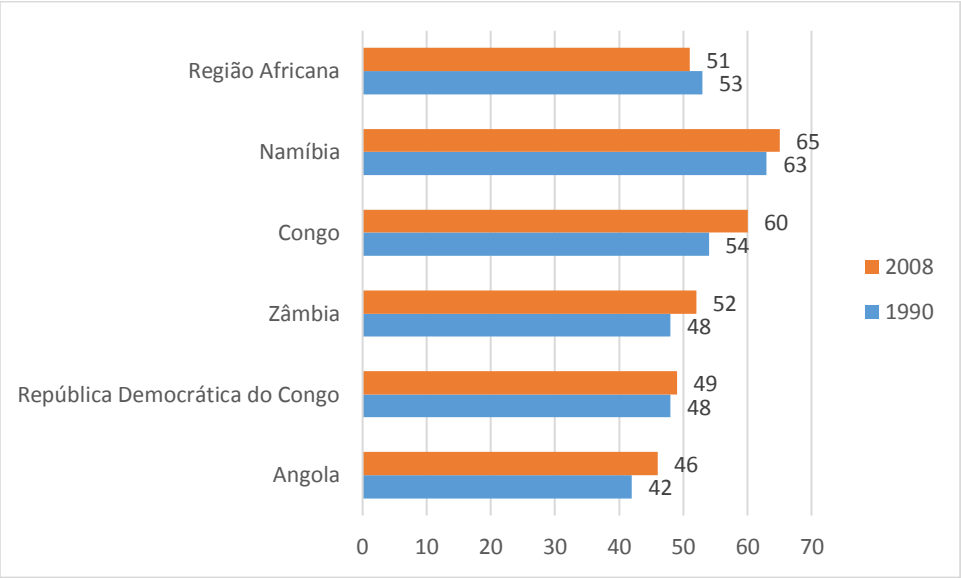
Desta forma, surge o pensamento que vincula a realidade de saúde à prática social, pois a compreensão e explicação de um fenómeno dependem da descoberta das relações e conexões que lhe são intrínsecas, que o formam e que o inserem na totalidade (Pilon, 1986, citado por Gesser, 2005), como proposto nos objetivos deste estudo. Assim sendo, o pensamento da saúde não pode ser desvinculado do todo, que é a realidade social da comunidade em estudo.

Portanto, a esperança de vida em qualquer região ou sociedade reflete o nível de desenvolvimento socioeconómico e demográfico do mesmo.

### 2.1.1 A esperança de vida em Angola

Falar-se da esperança de vida em Angola torna-se difícil tendo em conta a falta ou exiguidade de dados. Como se sabe, Angola esteve quase três décadas de conflito armado o que de certa medida fez com que as autoridades de direito não pudessem planificar ou alocar recursos para a existência de um Instituto Nacional de Estatística actuante, na medida em que no momento não era prioridade das autoridades governamentais a recolha de tais dados, ou a realização de estudos tendentes a melhoria das condições socioeconómicas das populações. Mas, apesar destas anomalias o país sempre contou com a presença de organizações não-governamentais tais como a Organização Mundial da Saúde, o Fundo das Nações Unidas para a População, a organização Médicos Sem Fronteira entre outros que de certo modo iam auxiliando as autoridades na recolha de dados, o que possibilitava de certo modo estimar alguns paramétros como é o caso da esperança de vida em média ao nascer em Angola. Assim, segundo dados recolhidos das páginas principais destas organizações, nota-se que a estimativa da esperança de vida fora feita até por volta de 2012 no geral, não se considerando a distribuição demográfica da população em diferentes regiões do país. Assim, e tendo em conta os dados recolhidos, apresentamos desde já a evolução da esperança de vida em angola e países vizinhos.

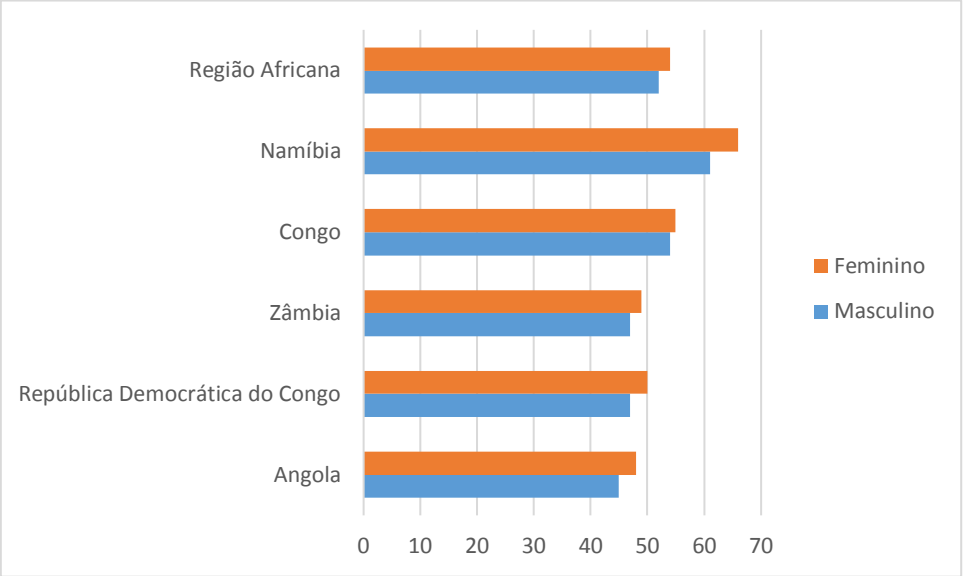
**Gráfico 2.1** Esperança de vida ao nascer em anos, em Angola e países vizinhos em 1990 e 2008.




**Fonte:** (OMS, 2010).

Do Gráfico 2.1 podemos notar que Angola no conjunto dos países vizinhos da África Austral é o país com menor índice de esperança de vida ao nascer em 1990 e em 2008 que corresponde respectivamente a 42 e 46 anos em média. É de salientar que apesar do índice de esperança de vida ao nascer ser baixo, podemos verificar que a mesma apresenta uma evolução positiva, isto é, de 1990 a 2008 houve um incremento de 4 anos.

**Gráfico 1.2** Esperança de vida ao nascer, em Angola e países vizinhos, por sexo, em 2008




**Fonte:** OMS (2010).

Da Figura 2.2 podemos notar que a esperança de vida em Angola ao nascer em 2008 por sexo foi dos mais baixo da região austral. Um outro padrão é o facto de as mulheres possuírem uma esperança de vida maior que a dos homens. Este fenómeno deve-se sobre tudo a forma como a sociedade encara o género, ou seja, na maior parte das sociedades ainda impera o machismo, o que torna o homem mais activo na sociedade. Para além disso, é de notar o carácter aventureiro dos homens em si. Uma outra causa que estará por detrás deste fenómeno, tem a ver com a taxa de natalidade que é maior para o género feminino em Angola e não só, segundo dados da OMS.

#### *2.1.1.1 Características da População*

- Dinâmica de distribuição da População

Segundo o Inquérito de Indicadores Básicos de Bem-Estar | QUIBB de 2011, entre 2008 e 2011 a população de Angola passou de 16.367.880 para 17.992.033 de habitantes com uma taxa de crescimento de 9,92%. As estimativas para o ano de 2012 apontavam para uma população de 18.576.568 habitantes, sendo 8.999.074 do sexo masculino e 9.577.494 do sexo feminino (INE, , 2013).

**Tabela 1.2** - Distribuição da população de 2008-2012 e taxa de crescimento

| Ano | Total | Homens | Mulheres | Taxa de cresccimento |
|---|---|---|---|---|
| 2008 | 16 367 880 | 7 898 969 | 8 468 911 | 3,01 |
| 2009 | 16 888 858 | 8 158 550 | 8 730 308 | 3,18 |
| 2010 | 17 429 637 | 8 427 802 | 9 001 835 | 3,20 |
| 2011 | 17 992 033 | 8 707 868 | 9 284 165 | 3,23 |
| 2012 | 18 576 568 | 8 999 074 | 9 577 494 | 3,25 |

Fonte: INE (Projecção da população 2009-2015)

Os dados sobre as projecções da população 2009-2015 indicam que no ano de 2011 63% da população de Angola residia em apenas cinco províncias do país. Em primeiro lugar aparece a capital do país, Luanda, com 28%, e cerca de 34% na região Centro-Sul, nomeadamente Kuanza-Sul (6,7%), Huíla (10,1%), Benguela (9,6%) e Huambo (8,0%). As províncias menos populosas encontram-se na região Norte (INE, , 2013).

**Tabela 1.3** - Distribuição da população por província e por $Km^2$ (2011)

| Província | População Total | Distribuição | Área | Densidade demográfica |
|---|---|---|---|---|
| Cabinda | 394 620 | 2,2 | 7 270 | 54,3 |
| Zaíre | 354 627 | 2,0 | 40 130 | 8,8 |
| Uíge | 945 196 | 5,3 | 58 698 | 16,1 |
| Luanda | 5 046 323 | 28,0 | 24 651 | 204,7 |
| Kuanza Norte | 330 979 | 1,8 | 24 190 | 13,7 |
| Kuanza Sul | 1 198 758 | 6,7 | 55 660 | 21,5 |
| Malange | 653 618 | 3,6 | 97 602 | 6,7 |
| Lunda-Norte | 684 417 | 3,8 | 102 783 | 6,7 |
| Benguela | 1 726 057 | 9,6 | 31 788 | 54,3 |
| Huambo | 1 443 388 | 8,0 | 34 274 | 42,1 |
| Bié | 1 003 042 | 5,6 | 70 314 | 14,3 |
| Moxico | 493 019 | 2,7 | 223 023 | 2,2 |
| Kuando-Kubango | 353 619 | 2,0 | 199 049 | 1,8 |
| Namibe | 324 673 | 1,8 | 58 137 | 5,6 |
| Huíla | 1 818 382 | 10,1 | 75 002 | 24,2 |
| Cunene | 570 918 | 3,2 | 89 342 | 6,4 |
| Lunda-Sul | 342 063 | 1,9 | 45 649 | 7,5 |
| Bengo | 308 333 | 1,7 | 31 371 | 9,8 |
| **Total** | **17 992 033** | **100** | **1 246 700** | **14,4** |

**Fonte**: INE (2011)

- Estrutura etária e sexo

O estudo sobre a estrutura da população por sexo e grupos etários permite analisar as tendências de fecundidade, mortalidade e migrações.

A estrutura da população pode ser resumida em três grandes grupos etários: 0-14 anos (grupo dos mais jovens), 15-64 anos (grupo dos potencialmente activos) e 65 ou mais anos (grupo dos idosos) (INE, Inquérito de Indicadores Básicos de Bem-Estar | QUIBB, 2013).

T**abela 1.4** - Distribuição dos principais grupos etários (1996, 2001,2008-2010 e 2011)

| Ano | 0-14 | 15-64 | 65 ou mais |
|---|---|---|---|
| 1996 – MICS I | 50,2 | 48,7 | 1,3 |
| 2001 – MICS II | 49,0 | 51,0 | 2,0 |
| 2008 – 2009 IBEP | 47,7 | 49,7 | 2,5 |
| 2011 – QUIBB | 47,8 | 49,7 | 2,5 |

**Fonte**: INE (2013)

A pirâmide etária mostra a estrutura da população por grupos etários, para homens e mulheres, separadamente. Os grupos etários mais jovens encontram-se na parte inferior e na parte superior encontram-se os grupos etários com mais idade. Assim, a pirâmide etária de Angola é caracterizada por uma base muito larga e um estreitamento no topo, típica dos países em desenvolvimento, reflectindo uma alta taxa de fecundidade (6,4 filhos), uma alta taxa de mortalidade (191,1) e uma baixa esperança de vida (42 anos) (INE, 2013).

A base da pirâmide é larga, mostrando que 47,8% da população pertence ao grupo etário dos 0-14 anos, 49,7% ao grupo etário com 15-64 anos e apenas 2,5% ao grupo dos idosos (65 anos ou mais). Cerca de 52% da população é do sexo feminino e 48% do sexo masculino (INE, Inquérito de Indicadores Básicos de Bem-Estar | QUIBB , 2013).

**Gráfico 1.3** Distribuição da população, segundo o sexo e a idade em 2011


**Fonte: INE** (2013).

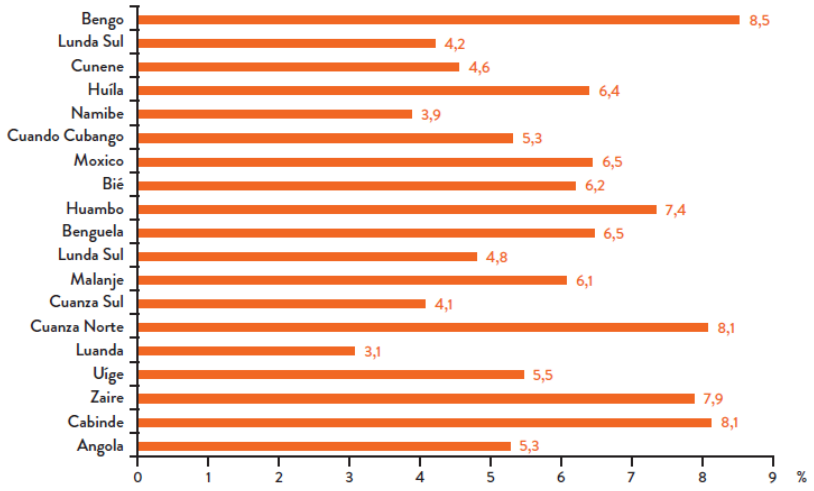
- Índice de Envelhecimento

Em Angola a proporção da população idosa é de 2,5%. O índice de envelhecimento reflecte este facto.

O índice de envelhecimento indica a relação entre a população idosa (65 anos ou mais) e a população mais jovem (0-14 anos). Os resultados do QUIBB – 2011 apontam para um índice de envelhecimento de 5%, isto é, para cada 100 pessoas com 0-14 anos de idade em Angola, existem apenas 5 idosos, reflectindo o predomínio da população jovem sobre a população idosa. O baixo índice de envelhecimento representa uma preocupação socioeconómica na avaliação das políticas sociais e de sustentabilidade.

O agravamento do não envelhecimento da população é quase comum em todas as províncias do país. Os menores índices de envelhecimento foram registados na província de Luanda e do Namibe com 3,1% e 3,9%, respectivamente.

**Gráfico 1.4** Índice de envelhecimento, segundo a província



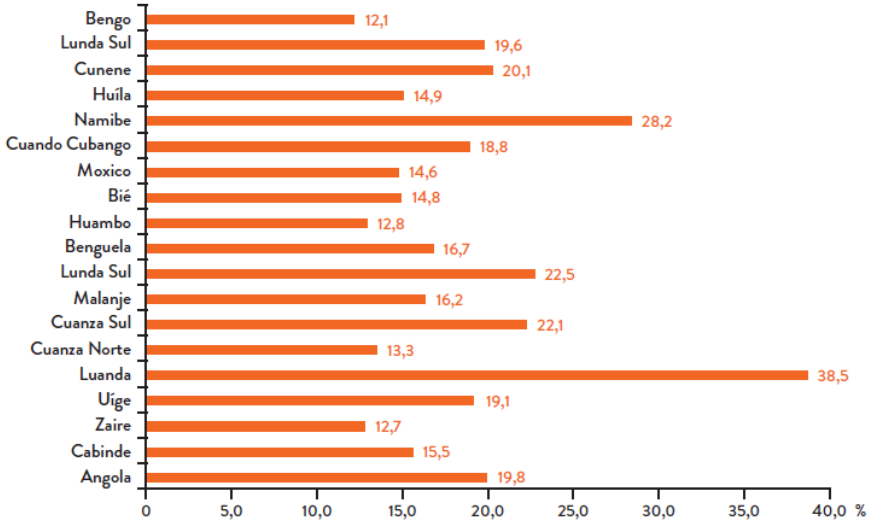

**Fonte**: INE (2013)

- Índice de Sustentabilidade Potencial

O índice de sustentabilidade é a relação existente entre a população em idade activa (15-64 anos) e a população idosa (65 anos ou mais). O índice de sustentabilidade é outro indicador que possibilita a avaliação sobre o esforço que a população idosa exerce sobre a população em idade activa. A nível nacional, o índice de sustentabilidade potencial é de 19,8 o que significa que em Angola há 19,8 pessoas em idade activa por cada pessoa com 65 anos ou mais. As províncias com maior índice de sustentabilidade são as

províncias de Luanda (38,5), Namibe (28,2) e Lunda Norte (22,5) (INE, Inquérito de Indicadores Básicos de Bem-Estar | QUIBB , 2013).

**Gráfico 1.5** Índice de sustentabilidade potencial, segundo a província




**Fonte**: INE (2013)

### 2.1.2 Cálculo da Esperança de Vida

Para o cálculo da esperança de vida ao nascer leva-se em consideração não apenas os riscos de morte na primeira idade – mortalidade infantil -, mas para todo o histórico de mortalidade de crianças, adolescentes, jovens, adultos e idosos. Sendo uma síntese da mortalidade ao longo de todo o ciclo de vida dos indivíduos, a esperança de vida é o indicador empregado para medir as dimensões humanas no índice de desenvolvimento, ou seja, direito a uma vida longa e saudável. Isso porque em cada um dos grupos etários os indivíduos estão sujeitos a diferentes riscos de mortalidade, estabelecendo distintas causas principais de mortalidade (PNUD, 2006).

O cálculo da esperança de vida serve para expressar a probabilidade de tempo médio de vida da população, representando uma medida sintética da mortalidade, não estando afetada pelos efeitos da estrutura etária da população (OPAS, 2004 *apud* Gesser, 2005).

Através deste índice torna-se possível analisar variações geográficas e temporais da esperança de vida da população, avaliando o nível de vida e saúde das pessoas e subsidiando os processos de planeamento, gestão e avaliação de políticas de saúde e previdenciárias entre outras (OPAS, 2004 *apud* Gesser, 2005).

### 2.1.3 Variáveis que afectam a esperança de vida

Existe um grande número de variáveis que podem afectar ou influenciar os padrões de esperança de vida de uma determinada população numa determinada região. Aqui podemos citar algumas, tais como o acesso aos serviços sociais básicos (escolas e hospitais) e o grau de pobreza. Podem ser de natureza económica, como o rendimento *per capita*, entre outras. Estas são normalmente consideradas em alguns estudos.

#### *2.1.3.1 Taxa de mortalidade por VIH/SIDA*

Em Angola, a recolha, a organização, o tratamento e a divulgação dos dados estatísticos relevantes é da responsabilidade do INE. Como se sabe Angola situa-se na África Subsaariana, região do globo com maiores índices de sero prevalência o VIH-SIDA e consequentemente de mortalidade pela SIDA. Segundo os relatórios da OMS, em 2008 registaram-se 1302668 óbitos que tiveram como causa o VIH-SIDA, sendo 593601 masculinos e 709067 femininos, em todas faixas etárias com principal destaque para as crianças dos 0 aos 4 anos de idade correspondendo a 88875 rapazes e 86475 meninas. Portanto, são números bastante alarmantes. O VIH-SIDA é já considerado como sendo uma das principais causas de mortalidade em África no geral e em particular na África Subsaariana. Em Angola, infelizmente os números são assustadores e apesar dos esforços que o governo vai levando acabo, tem havido incremento de casos de sero prevalência e de morte por VIH-SIDA.

Segundo o Anuário das Estatísticas sociais publicadas pelo INE, as Unidades de Saúde com Serviço de Aconselhamento e Testagem para o VIH-SIDA em 2007 foi de 154, sendo que 53% das unidades correspondia a mulheres gestantes e 47% a outros segmentos da população. Por outro lado, existe uma proporcionalidade relativa da distribuição das unidades de saúde, com serviço de aconselhamento e testagem para o VIH-SIDA, quanto aos casos positivos do VIH-SIDA segundo o grupo de idade em 2007 por províncias. Portanto, um outro dado que devemos reter é o facto das províncias de Luanda, Cunene, Cabinda e Huíla estarem no topo em termos de seroprevalência do VIH-SIDA e óbitos por VIH-SIDA. A Figura 6 mostra os óbitos relativos a algumas doenças transmissíveis por província em 2005.

**Tabela 1.5** Óbitos por Doenças Transmissíveis por Província, 2005 - 2007

| Província | 2005 | | | | |
|---|---|---|---|---|---|
| | Meningilte | Tuberculose pulmonar | Sida | Má Nutrição Aguda | Outras |
| Cabinda | 12 | 10 | nd | 1 | nd |
| Zaire | nd | nd | 1 | 1 | 25 |
| Uige | 6 | 29 | 3 | 8 | 20 |
| Luanda | 332 | 785 | 316 | 467 | nd |
| Kuanza-Norte | nd | 3 | nd | 1 | nd |
| Kuanza-Sul | 3 | 17 | nd | 26 | 3 |
| Malange | 4 | nd | nd | 1 | nd |
| Lunda-Norte | nd | 1 | nd | 4 | 1 |
| Benguela | 5 | 46 | 1 | 79 | 6 |
| Huambo | 10 | 86 | 4 | 29 | 5 |
| Bié | 15 | 49 | 2 | 58 | 6 |
| Moxico | 1 | 1 | nd | 5 | nd |
| Kuando Kubango | 1 | 34 | 2 | 7 | 14 |
| Namibe | nd | nd | nd | n | nd |
| Huila | 6 | nd | 7 | 9 | 1 |
| Cunene | 8 | 40 | 49 | nd | nd |
| Lunda-Sul | 1 | 8 | nd | nd | nd |
| Bengo | nd | nd | nd | nd | nd |

**Fonte:** INE (2011)

Conforme se pode notar, registaram-se 385 mortes por VIH-SIDA. Uma questão interessante neste quadro é o facto de não existirem registos de 9 províncias em pleno tempo de paz e numa fase em que se preparavam as eleições de 2008. Quero dizer que apesar de serem informações dos órgãos oficiais, não justificava, na altura, a falta de dados nestas províncias.

Segundo a OMS, a estimativa da seroprevalência do VIH-SIDA em Angola no período de 2001-2011 foi de 1165 por ano por cada 100.000 habitantes. Já a taxa de mortalidade por VIH-SIDA foi de 57 em 2001 e 59 em 2011 por cada 100.000 habitantes e os enfermos foi de 142 e 119 por cada 100.000 habitantes em 2001 e 2011, respectivamente.

Portanto, esta exiguidade de dados pode levar-nos a concluir que os números seriam muito maiores do que os publicados.

#### *2.1.3.2 Taxa de mortalidade por malária*

Segundo o INE (2011) na sua publicação Anuário de Estatística Sociais de 2009, nos anos de 2005 - 2007 a principal doença transmissível foi a malária com um total de 2.324.323 e 2.736.124 de casos em 2005 e 2007, respectivamente. Sendo que as províncias mais populosas são aquelas que apresentaram maior incidência

como são os casos de Luanda e Huambo. Por outro lado, houve um acréscimo de casos em 2007 se compararmos com o ano de 2005. Além disso, a malaria, dentre as várias doenças transmissíveis, foi a epidemia que causou mais morte ou óbitos de 2005 a 2007 com um total de óbitos de 13804 em 2005, 10352 em 2006 e 9503 em 2007. Como era de se esperar, as províncias com mais população e com maior incidência da malaria, foram as que apresentaram maiores casos de óbitos. Aqui, há a ressaltar o facto de ter existido algum decréscimo entre 2005 e 2007. Ainda segundo a OMS (2010), Angola foi um dos países na região com maiores casos notificados num total de 3,43 milhões. Um facto curioso, é a Vizinha Namíbia ter erradicado a mesma doença. Para mais detalhes verificar a Tabela 2.6.

**Tabela 1.6** Óbitos por Doenças Transmissíveis mais Frequentes por Província, 2005 - 2007

| Província | Malária | | | Diarreia* | | | Doenças Respiratórias** | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2005 | 2006 | 2007 | 2005 | 2006 | 2007 | 2005 | 2006 | 2007 |
| **Angola** | **13 804** | **10 352** | **9 503** | **1 992** | **1 330** | **1 412** | **1 766** | **1 528** | **1 682** |
| Cabinda | 132 | 175 | 50 | 4 | 7 | 4 | 8 | 17 | 2 |
| Zaire | 125 | 240 | 126 | 1 | nd | 7 | nd | 2 | 6 |
| Uíge | 1 175 | 609 | 719 | 114 | 89 | 54 | 138 | 112 | 72 |
| Luanda | 2 962 | 1 892 | 2 085 | 719 | 545 | 596 | 305 | 724 | 916 |
| Kuanza-Norte | 443 | 466 | 547 | 31 | 1 | 7 | 16 | 4 | 12 |
| Kuanza-Sul | 1 044 | 1 056 | 663 | 100 | 69 | 123 | 59 | 60 | 65 |
| Malange | 223 | 103 | 131 | 11 | 11 | 9 | 6 | 29 | 20 |
| Lunda-Norte | 460 | 1 343 | 1 429 | 154 | 78 | 78 | 181 | 75 | 95 |
| Benguela | 2 231 | 1 626 | 1 433 | 288 | 249 | 275 | 150 | 230 | 232 |
| Huambo | 1 981 | 113 | 102 | 122 | 13 | 5 | 16 | 3 | 4 |
| Bié | 717 | 749 | 526 | 31 | 69 | 85 | 426 | 58 | 72 |
| Moxico | 296 | 90 | 141 | 83 | 25 | 20 | 309 | 5 | 19 |
| Kuando Kubango | 602 | 827 | 599 | 192 | 93 | 78 | 53 | 124 | 53 |
| Namibe | 58 | 76 | 9 | 1 | 14 | nd | nd | 17 | 1 |
| Huila | 693 | 504 | 253 | 103 | 11 | 8 | 60 | nd | 62 |
| Cunene | 317 | 343 | 526 | 27 | 31 | 40 | 32 | 25 | 27 |
| Lunda-Sul | 52 | 96 | 66 | 6 | 8 | 3 | 3 | 18 | 15 |
| Bengo | 293 | 44 | 98 | 5 | 17 | 20 | 4 | 25 | 9 |

**Fonte:** INE (Inquérito Integrado sobre o Bem-Estar da População | IBEP, 2010)

A malária (paludismo) em Angola ainda é a primeira causa de morte e de doença, bem como do absentismo laboral e escolar. Representa cerca de 35% da procura de cuidados curativos, 20% dos internamentos hospitalares, 40% das mortes perinatais e 25% da mortalidade materna (Programa Nacional do Controlo da Malária, 2010). A malária tem não só um impacto negativo sobre a saúde das populações como também sobre o

desenvolvimento social destas, tornando-as mais pobres (INE, Inquérito de Indicadores Básicos de Bem-Estar | QUIBB , 2013).

O uso de redes mosquiteiras tratadas com insecticida (REMTI) pela população é considerado como um método extremamente eficaz e económico na prevenção contra a malária. Por este motivo, o Governo de Angola, através do Programa Nacional de Combate à Malária, tem intensificado a promoção e distribuição das redes mosquiteiras. Tal medida visa reduzir a morbilidade e mortalidade devidas à malária, principalmente nas crianças menores de cinco anos de idade e mulheres grávidas. Os dados revelam que a taxa de utilização de redes mosquiteiras é baixa, apenas cerca de um quarto da população (24,4%) dormiu debaixo de uma rede mosquiteira tratada com insecticida na noite anterior ao inquérito. Esta proporção apresenta variações significativas entre o meio urbano (27,1%) e rural (21,1%) e entre o quintil mais baixo (12,3%) e o quintil mais elevado (28,8%). De referir ainda que a população mais rica tem duas vezes mais hipóteses de dormir debaixo de uma rede mosquiteira tratada do que a mais pobre (INE, Inquérito de Indicadores Básicos de Bem-Estar | QUIBB , 2013).

No total 38% das crianças menores de cinco anos dormiram debaixo de uma rede mosquiteira de qualquer tipo e cerca de um terço (30,3%) dormiram debaixo de uma rede mosquiteira tratada com insecticida. Os resultados da Tabela 2.7 mostram as diferenças no uso da rede mosquiteira tratada com insecticida, segundo a área de residência (35,7% no meio urbano contra 23,8% no meio rural), mas as grandes disparidades verificam-se nos níveis de pobreza e níveis de escolaridade do chefe do agregado familiar. Das crianças do quintil mais baixo, 15,7% dormiram debaixo de uma rede mosquiteira tratada com insecticida, contrastando com os 40% do quintil mais elevado.

A situação é similar para o nível de escolaridade do chefe do agregado familiar, 21,7% das crianças cujo chefe do agregado não tem nenhum nível de escolaridade dormiram debaixo de uma rede mosquiteira tratada com insecticida, contrastando com os 41,6% das crianças cujo chefe do agregado tem o ensino secundário ou mais (quase o dobro).

**Tabela 1.7** Percentagem de crianças com 0-4 anos que dormiram debaixo de uma rede mosquiteira durante a noite anterior ao inquérito, segundo o tipo de rede

| Caracterização | Tratada com insecticida | Não tratada | Não sabe se era tratada | Qualquer tipo de rede | Número de pessoas |
|---|---|---|---|---|---|
| Angola | 30,3 | 5,9 | 1,8 | 38,0 | 8 197 |
| Urbana | 35,7 | 4,2 | 0,8 | 40,7 | 4 778 |
| Rural | 23,8 | 7,8 | 3,0 | 34,6 | 3 419 |
| Nenhum nível de ensino | 21,7 | 5,9 | 2,6 | 30,2 | 3 664 |
| Ensino primário | 34,3 | 6,3 | 1,2 | 41,8 | 2 211 |
| Ensino secundário ou mais | 41,6 | 5,3 | 1,1 | 48,1 | 2 320 |
| Primeiro | 15,7 | 5,8 | 4,0 | 25,5 | 1 676 |
| Segundo | 24,5 | 8,2 | 1,0 | 33,8 | 1 744 |
| Terceiro | 36,0 | 7,3 | 2,2 | 45,4 | 1 815 |
| Quarto | 36,1 | 4,9 | 1,2 | 42,3 | 1 700 |
| Quinto | 40,0 | 3,2 | 0,9 | 44,2 | 1 260 |

**Fonte:** INE, 2013).

Perante tais factos, podemos apontar como principais causas destes números alarmantes o fraco sistema de saúde de Angola, a falta de saneamento básico e o elevado nível de analfabetismo. As consequências são conhecidas e podemos apontar a existência de um deficiente programa de combate ao plasmódio que passaria pela eliminação dos principais vectores do mesmo. Por outro lado, nota-se que em vez de se combater a reprodução do plasmódio, procura-se combater o efeito com a distribuição de mosquiteiros à população, ignorando neste processo a equidade. Portanto, os mosquiteiros, em alguns casos, não chegam à população mais vulnerável e se chegam não estão impregnados com insecticidas.

#### *2.1.3.3 Taxa de mortalidade infantil (mortes/1.000 nascimentos normais)*

A taxa de mortalidade infantil é definida como sendo o número de óbitos de crianças com menos de 1 ano de idade ocorridos durante os últimos 12 meses anteriores ao Inquérito (habitualmente expressa por 1.000 nados vivos). A taxa de mortalidade infantil e a taxa de mortalidade em crianças menores de 5 anos apresentam comportamentos similares quando analisadas segundo o estado de pobreza. Crianças pobres têm taxas de mortalidade mais elevadas que crianças não-pobres durante os primeiros doze meses de vida (129 e 106 mortes por cada 1.000 crianças nascidas vivas, respectivamente) e antes de completarem cinco anos (218 e 176 mortes por cada 1.000 crianças nascidas vivas, respectivamente). Três resultados interessantes foram observados por áreas de residência (ver Tabela 2.8). Em primeiro lugar, ambas as taxas de mortalidade são mais elevadas em áreas rurais dos que nas cidades. Em segundo lugar, a diferença nas probabilidades de crianças pobres morrerem em áreas rurais em comparação com as cidades é relativamente baixa, enquanto que entre crianças não-pobres esta diferença é significativamente alta. Em terceiro lugar, nas áreas urbanas, as taxas de mortalidade entre crianças pobres são mais altas do que as observadas entre crianças não pobres, ao passo que, nas áreas rurais, as taxas de mortalidade entre crianças pobres são mais baixas que as observadas entre crianças não-pobres. O nível de escolaridade da mãe mostra diferentes associações com as taxas de mortalidade segundo o estado de pobreza. No caso das crianças pobres, o nível de escolaridade da mãe não parece ter qualquer efeito sobre as taxas de mortalidade, um resultado que deve ser interpretado com extrema cautela porque pode estar, provavelmente, relacionado com tamanhos de amostra reduzidos. No caso das crianças não pobres, a probabilidade delas morrerem diminui com os níveis mais altos de escolaridade da mãe. Por exemplo, as taxas de mortalidade quando a mãe não tem qualquer nível de escolaridade são quase o dobro das taxas de mortalidade quando a mãe tem pelo menos o ensino secundário. (INE, 2013).

**Tabela 1.8** Taxa de mortalidade infantil e de menores de cinco anos por estado de pobreza

| | Angola | Pobre | Não Pobre |
|---|---|---|---|
| TMI (0-12 meses) | 114,8 | 129.2 | 106,2 |
| Urbana | 93,7 | 121,5 | 88,7 |
| Rural | 138,6 | 130,8 | 148,9 |
| | | | |
| TMM5 (0-4 anos) | 191,9 | 217,8 | 175,5 |
| Urbana | 152,3 | 204,3 | 143,2 |
| Rural | 234,2 | 220,4 | 251,6 |

**Fonte:** INE (2013).

Os dados do INE são confirmados pelos relatórios do INE onde a taxa de mortalidade é em média de 115,7 e para menos de 5 anos é em média de 193,5. Para mais detalhe veja-se a Tabela 2.9.

**Tabela 1.9** Taxa de Mortalidade Infantil em Angola por sexo entre 2008-2009

(‰)

| | Taxa de mortalidade infantil | Taxa de mortalidade de menores de 5 anos de idade |
|---|---|---|
| **Angola** | **115,7** | **193,5** |
| Rapaz | 125,2 | 204,4 |
| Rapariga | 106,2 | 183 |

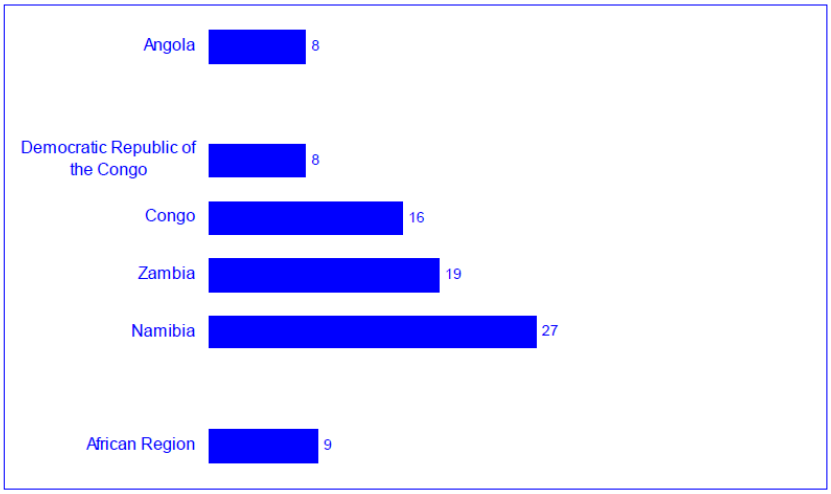
**Fonte**: INE (2010)

Portanto, apesar das causas apontadas pelas autoridades, pensamos que existem outras causas tais como, a falta de médicos nos principais centros de saúde e se existem não são especializados e em número bastante reduzido em relação a quantidade de utentes. Por outro lado, em alguns casos os dados estatísticos resultantes de campanha de vacinação são pouco fiáveis uma vez que são levados a cabo com pouca seriedade por quem de direito. Um exemplo deste facto é o que tem acontecido com as campanhas de vacinação contra a poliomielite que já se realiza há mais de duas décadas e os resultados deixam muito a desejar.

### *2.1.3.4 Camas hospitalares per capita (camas /1.000 habitantes)*

O número de camas hospitalares *per capita* é de extrema importância para o sistema de saúde de qualquer país. E angola não podia ser uma excepção. Infelizmente, existe uma exiguidade muito acentuada se compararmos o número de camas hospitalares por 10000 habitantes de Angola com os países vizinhos, entre 2000 e 2009. Os números são assustadores, conforme Figura nº 9. (WHO, Factsheets of Health Statistics, 2010)

**Gráfico 1.6** Número de camas hospitalares por 10.000 habitantes em Angola e Países vizinhos, 2008-2009




**Fonte**: WHO (2010)

Ainda segundo o INE (2011) no Anuário de Estatísticas Sociais de 2009, mostra que em 2005 o número de camas por cada 1000/habitantes foi de 0,77% , 1,04% em 2006 e 0.91% em 2007. Fica claro que houve um acréscimo entre 2005 e 2006 enquanto que entre 2006 e 2007 ocorreu um decréscimo. Ainda segundo esse relatório, nota-se que houve um investimento desigual relativamente a cada uma das províncias, sendo que a província com maior número de camas foi Luanda com 0,73 em 2005, 0,80 em 2006 e 0,96 em 2007. Seguindo-se Huambo 0,87% do número de camas de por 1000 habitantes em 2005, 0,31% em 2006 e 0,31% em 2007) e Benguela (com uma percentagem por cada 1000/habitantes do número de camas de 1,96 em 2005, 1,96 em 2006 e 2,19 em 2007).

Pelos factos apontados acima, é de notar que existiu na altura algum esforço por parte das autoridades governamentais no sentido de colmatar as insuficiências relativamente ao quesito camas hospitalares *per capita*. Mas, na nossa opinião as políticas nesse

sentido deveriam ser mais objectivas e dever-se-ia ter em atenção o conceito equidade, uma vez que a maior parte do investimento é feito apenas nas grandes superfícies urbanas em detrimento das rurais ou mais afastadas da capital, concentrando desta forma a maior parte das infraestruturas hospitalares nestes centros urbanos.

#### *2.1.3.5 Rendimento Familiar, Coeficiente Gini e Pobreza*

Segundo Ometto *et al.* (1995), citado por Gesser (2005), o padrão de vida de uma população depende de um conjunto de factores dentre os quais destaca-se a rendimento familiar calculada em termos reais, complementada ou não pela produção caseira de mercadorias e os benefícios sociais públicos e privados usufruídos por essa população. Logo, o padrão de vida de uma população vai depender dos espaços que se abrirem no mercado formal e informal de trabalho, dos salários e demais rendimentos nominais obtidos *versus* as taxas de inflação (e os preços relativos) verificados, além dos benefícios oferecidos pelos sectores públicos e privados.

Segundo Raphael (2000), citado por Gesser (2005), os padrões diferenciados de morbidade e mortalidade em grupos populacionais são determinados por múltiplos aspectos, entre eles destacam-se: a distribuição desigual dos factores de exposição e do acesso a bens e serviços de saúde, fragilidade das estruturas sociais de apoio à saúde e insuficiência de investimento em políticas sociais, especialmente em sociedades com grande nível de concentração de rendimento e baixo nível de coesão social.

Segundo Hart (1971), citado por Gesser (2005), o rendimento aparece como um factor que tem um grande impacto na saúde das pessoas, pois aqueles com menores condições socioeconômicas têm menor possibilidade de acesso aos serviços de saúde preventivos e curativos.

O Índice de Gini, que é uma das medidas mais conhecidas do grau de concentração do rendimento, apresenta uma variabilidade entre zero e um, sendo que este último valor corresponde à desigualdade máxima. É derivado através da curva de Lorenz, que é um gráfico que representa os percentuais acumulados de rendimento por decis da população. O índice é estimado conforme as diferenças entre as áreas delimitadas pela curva de Lorenz, o eixo horizontal e a reta de 45º (Hammond & Mccullagh, 1978, *apud* Gesser, 2005).

Em Angola, segundo o relatório do IBEP de 2008-2009, o nível de pobreza é muito alto e varia muito de com acordo a área de residência e por outro lado há uma diferença significativa entre os menores e os maiores quintis, havendo um incremento dos menores quanto aos maiores em termos de rendimento *per capita*. De forma geral em Angola, o rendimento média por pessoa para o 1º quintil foi de 1.414 Kwanzas (equivalente a 14 dólares americanos na altura). Enquanto que no 5º quintil o rendimento média mensal foi de 26.035 (o equivalente a 250 dólares americanos na altura). Analisando este dados, nota-se claramente que uma grande parte da população angolana vivia com menos de um dólar americano por dia. Em relação as áreas de residência o rendimento mensal em média por pessoa era maior nas zonas urbanas com um total de 11.077 Kwanzas equivalente ao rendimento *per capita* de 3,69 dólares americanos. Já para as zonas rurais, a situação era caótica uma vez que rendimentoo rendimento mensal em média por pessoa era de um total de 5.967 Kwanzas equivalente a um rendimento *per capita* de 1,989 dólares americanos. Relativamente ao rendimento médio mensal por pessoa por região, a capital figurava no topo, vindo a seguir as regiões Sul (9.187 Kwanzas), Centro Norte (6.972 Kwanzas), Norte (6.711 Kwanzas) e Centro Sul (7.435 Kwanzas). As regiões Este e Norte eram as mais pobres em termos de rendimento. Com rendimento médio por pessoa de 4.830 e 6.711 Kwanzas, respectivamente. Para mais esclarecimento consultar o Anexo 1. Relativamente a pobreza, em 2008-2009, segundo o relatório do IBEP os índices de pobreza podiam ser classificados em Incidência, Profundidade e Intensidade. Portanto, em termos de incidência 36,6% da população era pobre, em termos de Profundidade 12,7% era pobre e em termos de Intensidade 6,0% era pobre. Os índices de pobreza em termos de região eram menos profundos quanto à capital do país. Ainda temos a salientar que os índices de pobreza eram maiores nas zonas rurais tanto por Região quanto por região versus área de residência. Relativamente ao sexo as mulheres apresentavam índices de pobreza maiores. Em relação às idades os índices de pobreza eram maiores para as crianças e os jovens, conforme o quadro no Anexo 2.

Dos dados acima descritos, nota-se claramente que a população angolana e principalmente aquela das zonas rurais vivem no limiar da pobreza e falta-lhe quase tudo desde o acesso a serviços hospitalares dignos e de saneamento básico, infraestruturas sociais básicas como água potável, energia eléctrica, alimentação condigna, etc. Estas insuficiências ou falta de equidade na distribuição do bem comum

que se verifica na desigualdade social têm grande influência na diminuição da esperança de vida e no aumento das taxas de mortalidade. Alias, já no Brasil, Szwarcwald *et al.* (1999), citado por Gesser (2005), analisaram o impacto do coeficiente de Gini e da pobreza sobre a saúde das pessoas residentes na cidade do Rio de Janeiro, concluindo que existe uma correlação estatisticamente significante destas variáveis com as taxas de mortalidade e com a esperança de vida.

#### *2.1.3.6 O Acesso a água canalizada*

Apesar de todos os esforços para armazenar e diminuir o seu consumo, a água está se tornando, cada vez mais, um bem escasso, e sua qualidade se deteriora cada vez mais rápido (Freitas *et al*., 2001, *apud* Gesser, 2005).

Questões ambientais são fortemente relacionadas com as condições de vida e de saúde das populações. Regiões sem água canalizada e esgoto apresentam-se como regiões de risco para uma maior prevalência de doenças gastrointestinais e a cólera (Gerolomo & Penna, 2000, *apud* Gesser, 2005).

Condições ambientais precárias, abastecimento de água insuficiente e sistemas de esgoto inadequados são frequentemente citados como os maiores obstáculos para o controlo do desenvolvimento de surtos e epidemias (Medina, 1991, *apud* Gesser, 2005) como a de cólera (Witt; Reiff, 1991, *apud* Gesser, 2005).

Segundo a OMS no seu mais recente relatório de 2014, entre 1990 e 2000 apenas 42% da população tinham acesso a água potável, na primeira década do seculo 21 apenas 46% tinha acesso a água potável. Em 2012 52% da população já tinha acesso água potável (WHO, 2014). Nota-se que houve um incremento substancial de 6 pontos percentuais. Apesar disto, estes indicadores estão muito para além daquilo que deveria ser, uma vez que Angola possui recursos avultados que poderiam ser canalisados de uma forma mais aceitável. Não se compreende que em pleno século 21 as populações recorram as fontes não apropriadas de água para consumo como são os casos de riacho, rio ou charco e água da chuva ou das chipacas[1] (INE, 2010).

Segundo o relatório do IBEP de 2010, as principais fontes de água para beber classificavam-se em dois tipos:

1. Fontes apropriadas

[1] Lagos artificiais próprios para o bebedouro do gado

2. Fontes não apropriadas

Num total de 58.123 indivíduos inqueridos, 42% disse que usava fontes apropriadas de água para beber. O que quer dizer que mais de metade das pessoas inqueridas usavam fontes inapropriadas de água para beber. O que é lamentável tendo em conta as consequências resultantes de tal facto. Ainda segundo o mesmo relatório, quanto maior fossem o nível de escolaridade e o quintil de despesa melhor seria a fonte de água para beber.

#### *2.1.3.7 A Existência/inexistência de apoio médicm*

O artigo 21º da Constituição da República de 2010 sobre as tarefas do Estado Angolano na alínea f) diz "promover políticas que permitam tornar universais e gratuitos os cuidados primários de saúde" (Angola, 2010).

Ainda segundo a constituição da República no Artigo 77º sobre Saúde e protecção social no seu número 1 garante "0 Estado promove e garante as medidas necessárias para assegurar a todos o direito a assistência médica e sanitária, bem como o direito a assistência na infância, na maternidade, na invalidez, na deficiência, na velhice e em qualquer situação de incapacidade para o trabalho, nos termos da lei." (Angola, 2010).

Entretanto, a não obediência desse princípio dá-se em virtude da carência de recursos financeiros para o financiamento do setor, o que faz com que indivíduos com maior poder aquisitivo busquem os serviços privados de saúde como forma de garantir o acesso quando necessário. De acordo com o princípio de equidade, os serviços de saúde deveriam ser distribuídos segundo a necessidade de cuidados com a saúde, independentemente das características socioeconômicas individuais (Neri; Wagner, 2002, *apud* Gesser, 2005).

Muitas pesquisas em saúde (Campino et al., 1999; Le Grande, 1978; Almeida et al., 2000, *apud* Gesser, 2005, ), sugerem que a utilização dos serviços de saúde é bastante desigual entre classes socioeconómicas, favorecendo as camadas mais privilegiadas da população.

O último relatório da OMS espelha a falta de investimento tanto em recursos humanos quanto em infraestruturas e tecnologia. Por exemplo segundo o mesmo relatório em termos de densidade de profissionais de saúde por 10.000 habitantes existem 1,7 pediatras, 16,6 do pessoal de enfermagem e obstetrícia, menos de 0,05 de psiquiatras e

em relação a outro pessoal técnico infelizmente não existem dados ou os mesmos tendem a zero. Relativamente as infraestruturas e tecnologias há uma exiguidade quase que total quanto aos hospitais, às camas por hospitais e às camas psiquiátricas por cada 10.000 habitantes (WHO, 2014). Ainda em relação a este quesito, temos a salientar o facto de 0,4 unidades de tomografia computadorizada por cada 1.000.000 habitantes, menos de 0,05 de unidades de radioterapia por cada 1.000.000 de habitantes e 6,3 unidades de mamografia por 1.000.000 de mulheres com idade entre 60 e 69 anos.

O princípio da universalidade dos serviços de saúde garante que independentemente da condição social, ao cidadão lhe é garantido o essencial de acordo as suas necessidades – princípio de equidade.

Infelizmente, em Angola os serviços de saúde apesar de estarem constitucionalmente consagrados, não passam disto mesmo. Não existe confiança aos serviços públicos de saúde uma vez que os mesmos são bastante precários e deixam muito a desejar. Esta falta de confiança começa pelos políticos que preferem fazer qualquer tratamento fora de Angola. Por outra, os investimentos em termos de infraestrutura têm conhecido alguma melhoria, mas o fundamental que são os recursos humanos quase que pouco ou nada se tem feito. Tem-se recorrido a mão-de-obra estrangeira em particular a cubana no sentido de se ter um serviço de saúde com maior qualidade. Infelizmente essa aposta não tem resultado porque volta e meia os mesmos já estão a colaborar em clinicas privadas a partir das quais obtém rendimentos extras e porque a população mais "vulnerável" com alguma posse prefere ir a uma clínica ao invés de um hospital público.

#### *2.1.3.8 Acesso à Educação*

Dizia o líder carismático sul africano Nelson Mandela que: "A educação é a arma mais poderosa para mudar o mundo".

A educação deve ser entendida como uma forma de promoção do desenvolvimento do homem como indivíduo e como parte de um ambiente complexo (Pilon, 1986, *apud* Gesser, 2005), para que o mesmo consiga interagir com a sociedade aprimorando sua qualidade de vida. Portanto, a educação é um factor de integração social sem a qual o individuo fica totalmente excluído da sociedade, desconhecendo os seus direitos e seus deveres.

Segundo Buss (1999) citado por Gesser (2005) o acesso às informações é imprescindível para o exercício da cidadania, assim como iniciativas do poder público nos campos da educação e da comunicação em saúde. A educação, assim como a comunicação de massas, através de diversas médias, têm sido reconhecidas como ferramentas importantes que fazem parte da promoção da saúde de indivíduos e da comunidade.

Segundo Buss (1999) citado por Gesser (2005), a educação em saúde apresenta-se em dois focos: seja aquele mais concentrado na mudança de estilos de vida e fatores comportamentais do indivíduo, seja o que está centrado nas políticas públicas saudáveis e em ações comunitárias para se estabelecer prioridades, tomar decisões, planear e implementar estratégias visando melhorar os níveis de saúde populacionais.

Estas asserções de Buss mostram claramente o quão importante é a educação em questões que têm a ver com a saúde e o bem-estar. Em Angola, vários são os casos de relatos envolvendo morte por intoxicação. A causa muitas vezes apontada é o analfabetismo que faz com que haja uma comunicação deficiente entre o médico e o paciente, tendo como consequência a má interpretação da prescrição por parte deste último.

O padrão educacional é uma variável de grande relevância para as questões de saúde das populações, haja vista o impacto que este causa nos índices de mortalidade. Aycaguara e Macho (1990) conforme refere Gesser (2005), avaliaram, através de um estudo ecológico exploratório, o padrão de mortalidade infantil dos países do continente americano e relataram que o padrão educacional e a taxa de natalidade eram as variáveis que apresentavam maior grau de influência.

Quanto a Angola, os níveis de analfabetismo são ainda muito elevados apesar de algum esforço que se tem feito por parte dos organismos de tutela. Mas segundo o relatório do IBEP de 2010, em 2008-2009 a proporção de crianças com 3-5 anos de idade, segundo a frequência do ensino pré-escolar era de 9,3 num total 4.865 crianças inqueridas (INE, Inquérito Integrado sobre o Bem-Estar da População | IBEP, 2010). Aqui temos a referir que apesar de existirem varias crianças que procuram os serviços educacionais, verifica-se muita desistência principalmente nas zonas rurais. São vários os factores indicados pelos pais ou encarregados de educação. Uma das principais causas é a não existência

do serviço na localidade. Por outro lado, a área rural apresenta maior proporção, conforme Tabela 2.10.

**Tabela 1.10** Crianças com 3-5 anos de idade que não frequentam o ensino pré-escolar, segundo a razão.

| | A não frequentarem o ensino pré-escolar | Razão de não frequentarem o ensino pré-escolar | | | | | | | Total | Número de crianças com 3-5 anos |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Não existe serviço na localidade | Ainda é pequeno | Custa caro | Tem quem cuida dele(a) em casa | Fica muito longe | Não é importante | Outro | | |
| **Angola** | **90,7** | **38,1** | **38,0** | **15,1** | **3,5** | **2,0** | **0,7** | **2,7** | **100** | **4.865** |
| **Área** | | | | | | | | | | |
| Urbana | 89,0 | 25,3 | 34,1 | 28,0 | 5,7 | 2,2 | 0,9 | 3,7 | 100 | 2.441 |
| Rural | 92,7 | 52,2 | 42,3 | 0,8 | 1,0 | 1,7 | 0,4 | 1,5 | 100 | 2.424 |

**Fonte**: INE (2010).

Em relação ao ensino geral existem muitos alunos que atingiram ou frequentaram desde o ensino primário ao ensino secundário 2º ciclo, incluindo o ensino superior, considerando a amostra do levantamento feito pelo INE em 2008-2009. E de realçar que a maior proporção encontra-se no ensino primário e vai decrescendo até ao ensino superior como é natural. Um outro indicador é o facto de existir aqui alguma diferença significativa em relação a área de residência. Para os níveis primário e secundário do primeiro ciclo a área rural está no topo em relação a urbana. Já nos níveis mais acima ela fica abaixo da urbana. Um outro padrão é a existência de um certo equilíbrio para os níveis primários e secundários 1º ciclo entre todas as províncias. Mas em relação aos níveis secundário 2º ciclo e superior existe uma disparidade considerável, conforme Figura 2.11.

**Tabela 2.1.11** População com 18 ou mais anos de idade, segundo o nível de ensino atingido

| | Alguma vez frequentou a escola | Nível de ensino atingido | | | | | Total | Número de pessoas com 18 ou mais anos |
|---|---|---|---|---|---|---|---|---|
| | | Ensino primário | Ensino secundário 1° ciclo | Ensino secundário 2° ciclo | Ensino superior | Outros | | |
| **Angola** | **75,9** | **55,5** | **24,2** | **16,1** | **4,1** | **0,1** | **100** | **26.431** |
| **Área de residência** | | | | | | | | |
| Urbana | 88,6 | 41,4 | 29,6 | 22,7 | 6,2 | 0,1 | 100 | 14.019 |
| Rural | 59,9 | 82,2 | 13,9 | 3,6 | 0,1 | 0,1 | 100 | 12.412 |
| **Província** | | | | | | | | |
| Cabinda | 78,5 | 46,1 | 29,0 | 21,0 | 3,8 | 0,1 | 100 | 1.490 |
| Zaire | 80,1 | 59,0 | 24,4 | 15,3 | 0,4 | 0,9 | 100 | 1.340 |
| Uíge | 67,5 | 59,1 | 27,1 | 11,4 | 2,3 | 0,2 | 100 | 1.518 |
| Luanda | 93,2 | 34,3 | 32,9 | 24,7 | 8,0 | 0,0 | 100 | 3.516 |
| Kwanza Norte | 75,3 | 69,6 | 21,4 | 6,8 | 2,0 | 0,2 | 100 | 1.225 |
| Kwanza Sul | 63,7 | 79,7 | 15,3 | 4,7 | 0,4 | 0,0 | 100 | 1.265 |
| Malange | 62,4 | 68,8 | 17,5 | 12,7 | 0,6 | 0,5 | 100 | 1.336 |
| Lunda Norte | 55,4 | 75,3 | 14,5 | 8,4 | 1,7 | 0,1 | 100 | 1.376 |
| Benguela | 73,1 | 64,8 | 16,0 | 15,2 | 3,7 | 0,2 | 100 | 1.345 |
| Huambo | 66,1 | 75,3 | 13,2 | 8,7 | 2,6 | 0,1 | 100 | 1.304 |
| Bié | 66,3 | 81,4 | 13,6 | 5,0 | 0,0 | 0,0 | 100 | 1.127 |
| Moxico | 66,3 | 68,5 | 21,3 | 9,7 | 0,3 | 0,3 | 100 | 1.385 |
| Kuando Kubango | 54,1 | 66,9 | 21,2 | 9,6 | 0,8 | 1,4 | 100 | 1.177 |
| Namibe | 68,1 | 65,5 | 20,6 | 10,4 | 2,7 | 0,9 | 100 | 1.584 |
| Huíla | 71,7 | 64,3 | 21,2 | 13,4 | 1,2 | 0,0 | 100 | 1.300 |
| Cunene | 69,8 | 66,0 | 23,0 | 10,5 | 0,3 | 0,2 | 100 | 1.493 |
| Lunda Sul | 65,7 | 76,2 | 17,1 | 6,0 | 0,5 | 0,2 | 100 | 1.335 |
| Bengo | 76,5 | 77,8 | 17,3 | 4,4 | 0,4 | 0,1 | 100 | 1.315 |
| **Sexo** | | | | | | | | |
| Homens | 87,8 | 50,3 | 27,5 | 17,6 | 4,5 | 0,1 | 100 | 12.422 |
| Mulheres | 65,5 | 61,6 | 20,3 | 14,3 | 3,6 | 0,2 | 100 | 14.009 |
| **Idade** | | | | | | | | |
| 18 - 19 anos | 89,2 | 51,6 | 31,2 | 16,1 | 1,1 | 0,0 | 100 | 2.529 |
| 20 - 24 anos | 86,5 | 45,0 | 28,4 | 21,9 | 4,6 | 0,1 | 100 | 4.975 |
| 25 - 29 anos | 81,9 | 48,6 | 26,0 | 19,8 | 5,5 | 0,1 | 100 | 3.670 |
| 30 - 34 anos | 78,9 | 56,3 | 25,3 | 13,2 | 5,1 | 0,2 | 100 | 3.199 |
| 35 - 39 anos | 80,8 | 56,7 | 24,2 | 14,8 | 4,3 | 0,0 | 100 | 2.605 |
| 40 - 44 anos | 82,1 | 60,8 | 20,5 | 14,3 | 4,3 | 0,1 | 100 | 2.272 |
| 45 - 49 anos | 72,4 | 60,7 | 19,3 | 16,0 | 3,7 | 0,2 | 100 | 1.887 |
| 50 ou mais anos | 49,1 | 76,4 | 12,6 | 7,7 | 3,0 | 0,3 | 100 | 5.294 |

**Fonte:** INE (Inquérito Integrado sobre o Bem-Estar da População | IBEP, 2010)

Perante estes dados, fica claro que existem muitos problemas no âmbito da educação e os níveis de analfabetismo, ainda, são significativos, o que é nocivo à saúde e qualidade de vida aceitável em uma sociedade que cresce economicamente de forma assustadora. Na nossa opinião este crescimento deveria se repercutir na vida social dos cidadãos com maiores investimentos em infraestruturas de ensino, centros pré-escolares, formação continuada dos professores, melhores condições sociais e de trabalho dos mesmos, elaboração de currículos escolares actuais, etc.

### 2.1.4 Perspectivas Futuras em Esperança de Vida

Os fatores socioeconômicos são tão determinantes no cálculo da esperança de vida que os grupos pertencentes aos níveis socioeconômicos mais altos tem tido um incremento

mais significativo do que aqueles de nível socioeconômico mais baixo (Bremer et al., 2000, *apud* Gesser, 2011). Esta afirmação se faz presente em diversos estudos realizados com desenho ecológico, como, por exemplo, na República Tcheca (Richtarikova & Dzurova, 1992, *apud* Gesser, 2005) onde os fatores sociais, juntamente com a mortalidade infantil, apresentaram alto grau de significância com a esperança de vida ao nascer. Naquele estudo a variável socioeconômica que apresentou maior grau de associação com a esperança de vida foi a educação.

O aumento da esperança de vida promoveu o surgimento das mais variadas especulações a respeito do que acontecerá no século XXI. Para realmente entender o que poderá ocorrer com a esperança de vida da população. Segundo Olshansky et al. (2001) citado por Gesser (2005), compararam os dados demográficos de três países (Estados Unidos, França e Japão) no período de 1985 a 1995 e chegaram às seguintes conclusões:

1. A esperança de vida aumentou nos três países nesse período;
2. Quando a média de duração da vida de uma população se aproxima de 80 anos, os ganhos futuros em longevidade avançam em ritmo mais lento;
3. Nas próximas décadas, para que as esperanças de vida alcancem 85 anos, deverão ocorrer reduções muito drásticas nos índices de mortalidade em todas as faixas etárias. Para atingir esse objetivo entre as mulheres japonesas, o subgrupo que está mais próximo dele, a mortalidade geral deverá cair 20%; no caso das francesas, 26% e, no das americanas, mais do que 50%;
4. Projetando os números obtidos no período de 1985 a 1995, a esperança de vida dos franceses (homens e mulheres) só chegará aos 85 anos em 2033, a dos japoneses, em 2035 e a dos americanos, em 2182;
5. Para a esperança de vida ultrapassar 100 anos, mesmo em países com populações de grande longevidade, como França e Japão, será preciso eliminar todos os riscos de morte antes dos 85 anos.

Evidentemente que tais previsões estão baseadas em dados oriundos de países desenvolvidos e que tem parte de seus problemas sociais resolvidos. Assim sendo, antes de adotar-se essas previsões como verdadeiras e orientadoras de políticas públicas para Angola deve-se entender que as questões de natureza socioeconômica devam ser primeiramente equacionadas.

# 3 Regressão Linear

## 3.1 Introdução

A colecção de ferramentas estatísticas que são usadas para modelar e explorar relações entre variáveis que são relatadas em uma maneira não determinística é chamada análise de regressão. Tendo em conta que problemas deste tipo ocorrem tão frequentemente em muitos ramos da engenharia e da ciência, a análise de regressão é uma das ferramentas amplamente usadas. Neste estudo apenas trataremos da análise de regressão linear, cujo modelo pode ser dado por:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon\,, \tag{3.1}$$

onde $Y$ é a variável dependente ou variável resposta, os parâmetros $\beta_j$, $j = 0,1,\cdots,k$ são chamados de coeficientes de regressão, as variáveis $x_j$, $j = 1,\cdots,k$ são chamadas independentes ou regressoras enquanto que **ε** corresponde ao termo do erro aleatório.

## 3.2 Regressão Linear Simples

No modelo de regressão linear simples queremos relacionar duas variáveis através de um modelo linear, ou seja através da equação de uma recta dada por

$$Y = \beta_0 + \beta_1 x\,. \tag{3.2}$$

Os dados de que dispomos consistem no conjunto de $n$ pares de valores $(x_i, \mathrm{y}_i), i = 1,\cdots,n.$

Os valores $x_i$, $i = 1,...,n$, representam os valores da variável independente e são considerados determinísticos (pré-determinados à partida). Tal como foi referido anteriormente esta variável é habitualmente designada por regressora.

Os valores $Y_i$, $i = 1,...,n$, representam os valores da variável dependente e estes são considerados variáveis aleatórias.

Os coeficientes da recta, $\beta_0$ e $\beta_1$, são designados por coeficientes de regressão.

O modelo probabilístico será

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ i = 1,\ldots,n,$$

onde $\varepsilon_i$ representa o erro aleatório que provoca um afastamento na vertical dos pontos em relação à recta da equação $Y_i = \beta_0 + \beta_1 x_i$. De acordo com este modelo, se pretendermos que a recta «passe pelo meio dos pontos», é natural esperar que os erros $\varepsilon_i$ tenham média 0. Isso é o mesmo que dizer que a variável $Y_i$, quando pensada como função de $x_i$, tem média $\beta_0 + \beta_1 x_i$, ou seja, se observarmos vários valores $Y_i$ para a mesma abcissa $x_i$ eles devem dispor-se verticalmente em torno do ponto e a sua média (valor esperado) deve ser exactamente a ordenada do ponto da recta. Analiticamente isto significa que existe uma relação linear entre o valor esperado de $Y_i$ e o valor da regressora que lhe corresponde, $x_i$ (Hall, Neves, & Pereira, 2011). Tem-se portanto

$$E\left[Y_i\right] = \beta_0 + \beta_1 x_i \,. \tag{3.3}$$

Por forma a tornar mais simples a análise estatística deste tipo de modelos, pressupõe-se que os erros são independentes e identicamente distribuídos com distribuição normal de média 0 e variância $\sigma^2$, $\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$. Em geral não é necessário impor independência, basta impor que os erros sejam não – correlacionados (coeficiente de correlação linear nulo, $R = 0$). De qualquer forma, se impusermos que as variáveis sejam independentes então garantimos que não há nenhum tipo de relação entre elas pelo que garantimos que não são correlacionadas, independentemente da distribuição que tenham.

Ao assumir que os erros têm distribuição Normal, concluímos que também as observações $Y_i$ vão ter distribuição Normal já que $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, e a soma de uma Normal com uma constante tem distribuição Normal. Assim, para cada abcissa $x_i$,

$$Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), i = 1, \ldots, n. \tag{3.4}$$

### 3.2.1 Estimação dos parâmetros do modelo

A forma mais habitual de estimar os coeficientes de regressão baseia-se no método dos mínimos quadrados, que nas condições descritas anteriormente fornece os mesmos resultados que o método da máxima verosimilhança. O método dos mínimos quadrados fornece como solução a recta que minimiza a soma dos quadrados dos desvios das

observações $Y_i$ em relação aos seus valores esperados $E[Y_i] = \beta_0 + \beta_1 x_i + \varepsilon_i$, ou seja em relação à ordenada do ponto sobre a recta. Esses desvios são dados por

$$Y_i - E[Y_i] = \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 x_i = \varepsilon_i. \tag{3.5}$$

A soma dos quadrados dos desvios é então dada por

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{3.6}$$

Os estimadores dos mínimos quadrados de $\beta_0$ e $\beta_1$, designados por $\hat{\beta}_0$ e $\hat{\beta}_1$, devem satisfazer as seguintes igualdades

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= -2\sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\ \frac{\partial L}{\partial \beta_1} &= -2\sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \end{aligned} \tag{3.7}$$

Simplificando as duas equações, teremos:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{3.8}$$

As equações anteriores são chamadas de equações normais dos mínimos quadrados. A solução para as equações normais resulta nos estimadores dos mínimos quadrados dos parâmetros, $\hat{\beta}_0$ e $\hat{\beta}_1$.

As estimadores de mínimos quadrados dos parâmetros do modelo de regressão linear simples são

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\ \bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{3.9}$$

onde $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ e $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ .

A recta de regressão ajustada ou estimada será, portanto,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{3.10}$$

Note-se que cada par de observações satisfaz a relação

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, i = 1,2,\ldots,n. \tag{3.11}$$

onde $e_i = y_i - \hat{y}_i$ é chamado resíduo. O resíduo $e_i$ descreve o erro no ajuste do modelo da i-ésima observação, $y_i$, $i = 1,\ldots,n$.

Em termos de notação é conveniente ocasionalmente atribuir símbolos especiais no numerador e no denominador da fórmula do modelo como os que se seguem

$$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - \frac{\sum_{i=1}^{n} x_i}{n}$$

e

$$S_{xy} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}.$$

Utilizando esta notação simplificada podemos reescrever as expressões dos estimadores de mínimos quadrados dos coeficientes da recta de regressão da seguinte forma

$$\begin{cases} \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \\ \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \end{cases} . \tag{3.12}$$

Importa salientar que do ponto de vista estatístico o estimador de $\beta_1$ não é mais do que uma combinação linear dos valores $y_i$, $i = 1,\ldots,n$. Uma vez que os $x_i$, $i = 1,\ldots,n$, são considerados fixos, facilmente se conclui que $\hat{\beta}_1$ tem distribuição Normal pois é baseado na soma de Normais $(y_i)$ multiplicadas por constantes. De igual modo conclui-se que $\hat{\beta}_0$ também tem distribuição Normal. Resta saber quais os valores esperados e variâncias destas distribuições Normais (Hall, Neves, & Pereira, 2011).

#### *3.2.1.1 Propriedades dos estimadores dos mínimos quadrados*

As propriedades estatísticas dos estimadores dos mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser facilmente descritas. Assumiu-se anteriormente que o termo do erro $\varepsilon$ no modelo $Y = \beta_0 + \beta_1 x + \varepsilon$ é uma variável aleatória com média zero e variância $\sigma^2$. Desde que os valores de $x$ sejam fixados, $Y$ é uma variável aleatória com média $\mu = \beta_0 + \beta_1 x$ e variância $\sigma^2$. Portanto, os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ dependem das observações de $y$, assim, os estimadores dos mínimos quadrados dos coeficientes de regressão podem ser vistos como variáveis aleatórias. Poderemos investigar o viés e as propriedades da variância dos estimadores dos mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$.

Consideremos em primeiro lugar $\hat{\beta}_1$. Porque $\hat{\beta}_1$ é uma combinação linear das observações $Y_i$, podemos usar propriedades do valor esperado para mostrar que o valor esperado de $\hat{\beta}_1$ é, ver por exemplo Hall, Neves, & Pereira (2011, p. 104).

$$E\left(\hat{\beta}_1\right) = \beta_1 .$$

Assim, $\hat{\beta}_1$ é um estimador não enviesado de $\beta_1$.

Agora consideremos a variância de $\hat{\beta}_1$. Uma vez que assumimos que $V(\varepsilon_i) = \sigma^2$, segue-se que $V(Y_i) = \sigma^2$, porque $\hat{\beta}_1$ é uma combinação linear das observações $Y_i$. Resultando deste modo que (para um esclarecimento mais claro consultar a Secção 5-5 do livro de Montegomery & Runger, 2011).

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{3.13}$$

Para o intercepto, podemos mostrar de forma similar que (Hall, Neves, & Pereira, 2011, p. 277-278),

$$E(\hat{\beta}_0) = \beta_0 \text{ e } V(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]. \tag{3.14}$$

Assim, $\hat{\beta}_0$ é um estimador não enviesado do intercepto, $\beta_0$. A covariância das variáveis aleatórias $\hat{\beta}_0$ e $\hat{\beta}_1$ é não nula. Pode-se mostrar ainda que (Portal Action, 2015).

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}. \tag{3.15}$$

O estimador de $\sigma^2$ pode ser usado nas duas equações anteriores para fornecer estimativas da variância do declive e do intercepto. Chamamos raízes quadradas de estimadores da variância, resultantes, os erros padrões estimados do declive e do intercepto, respectivamente.

#### *3.2.1.2 Estimador de $\sigma^2$*

Existe um outro parâmetro desconhecido no modelo de regressão, trata-se de $\sigma^2$, que corresponde à variância do termo do erro. Os resíduos são usados para obter uma

estimativa para $\sigma^2$. A soma dos quadrados dos resíduos, frequentemente chamada soma dos quadrados dos erros, é dada por

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2. \tag{3.16}$$

Pode-se mostrar que o valor esperado da soma dos quadrados dos erros é $E\left(SS_E\right) = \left(n-2\right)\sigma^2$ (Hall, Neves, & Pereira, 2011, p. 279)

Portanto um estimador não enviesado (centrado) da $\sigma^2$ é

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}. \tag{3.17}$$

Calcular $SS_E$ usando a Equação (3.16) seria bastante tendencioso. Uma fórmula de cálculo mais conveniente pode ser obtida substituindo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ na Equação (3.16) e simplificando. A fórmula de cálculo resultante é

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}, \tag{3.18}$$

onde $SS_T = \sum_{i=1}^{n} \left( y_i - \overline{y} \right)^2 = \sum_{i=1} y_i^2 - n\overline{y}^2$ é o total da soma dos quadrados da variável resposta $y$.

### 3.2.2 Testes de Hipóteses na Regressão Linear Simples

Uma parte importante da avaliação da adequação de um modelo de regressão linear é testar hipóteses estatísticas acerca dos parâmetros do modelo e construir intervalos de confiança. Para testar hipóteses acerca do declive e do intercepto do modelo de regressão, devemos fazer uma suposição adicional que corresponde ao componente do erro no modelo, $\varepsilon$, ser normalmente distribuído. Assim, as suposições completas são

que os erros são independentes e normalmente distribuídos com média zero e variância $\sigma^2$, abreviado $\text{NID}\left(0,\sigma^2\right)$.

De referir ainda que na regressão linear simples o erro padrão estimado do declive e o erro padrão estimado do intercepto são respetivamente

$$se\left(\hat{\beta}_1\right)=\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{e} \quad se\left(\hat{\beta}_0\right)=\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}, \tag{3.19}$$

onde $\hat{\sigma}^2=\frac{SS_E}{n-2}$.

**Uso do teste** $t$-Student

Suponhamos que se deseja testar a hipótese de que o declive é igual a uma constante $\beta_{1,0}$

As hipóteses em abordagem são portanto

$$\begin{matrix} H_0:\beta_1=\beta_{1,0} \\ H_1:\beta_1\neq\beta_{1,0} \end{matrix}, \tag{3.20}$$

onde assumimos uma alternativa bilateral. Uma vez que os erros $\varepsilon_i$ são $\text{NID}\left(0,\sigma^2\right)$, segue-se directamente que as observações $Y_i$ são $\text{NID}\left(\beta_0+\beta_1 x_i,\sigma^2\right)$. Agora $\hat{\beta}_1$ é uma combinação linear das variáveis aleatórias normais independentes, e consequentemente,

$$\hat{\beta}_1 \sim N\left(\beta_1,\sigma^2/S_{xx}\right),$$

usando os resultados discutidos anteriormente. Além disso, $(n-2)\hat{\sigma}^2/\sigma^2$ tem uma distribuição qui-quadrado com $n-2$ graus de liberdade e $\hat{\beta}_1$ é independente de $\hat{\sigma}^2$. Como resultado destas propriedades, a estatística

$$T_0=\frac{\hat{\beta}_1-\beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}}, \tag{3.21}$$

segue uma distribuição $t$-Student com $n-2$ graus de liberdade, sob $H_0 : \beta_1 = \beta_{1,0}$. Rejeitamos $H_0 : \beta_1 = \beta_{1,0}$, para um nível de significância $\alpha$, se

$$|t_0| > t_{1-\alpha/2,n-2} \ ,$$

onde $t_0$ é o valor observado da estatística definida pela Equação (3.21) e $t_{1-\alpha/2,n-2}$ corresponde ao quantil $1-\alpha/2$ da distribuição *t*- Student com $n-2$ graus de liberdade. O denominador da equação é o erro padrão do declive, assim podemos escrever a estatística de teste como

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se\left(\hat{\beta}_1\right)} \tag{3.22}$$

Um procedimento similar pode ser usado para testar a hipótese acerca da intersecção. Para o teste

$$\begin{cases} H_0 : \beta_0 = \beta_{0,0} \\ H_1 : \beta_0 \neq \beta_{0,0} \end{cases} \tag{3.23}$$

devemos usar a estatística

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se\left(\hat{\beta}_0\right)}, \tag{3.24}$$

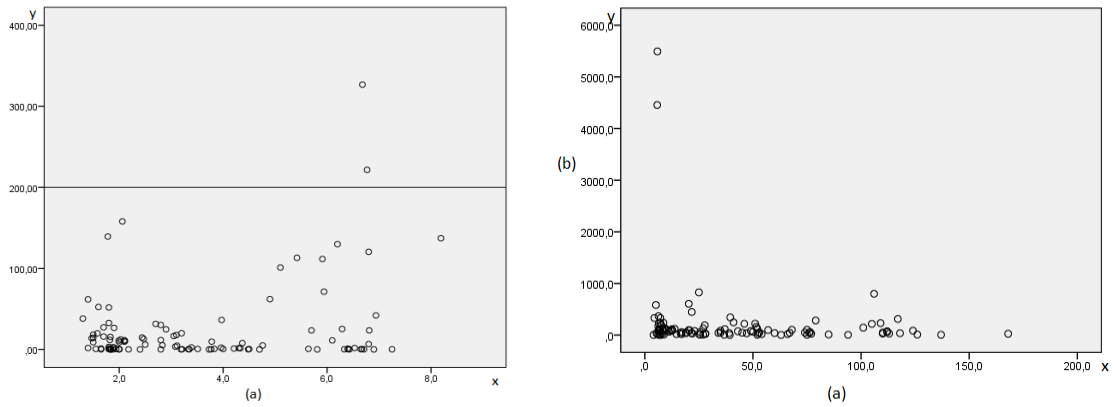
que também segue uma distribuição t-student com $n-2$ graus de liberdade. Deve-se rejeitar a hipótese nula se o valor observado desta estatística de teste, $t_0$, é tal que $|t_0| > t_{1-\alpha/2,n-2}$.

Um caso particular muito importante das hipóteses definidas na Equação (3.20) é

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}. \qquad (3.25)$$

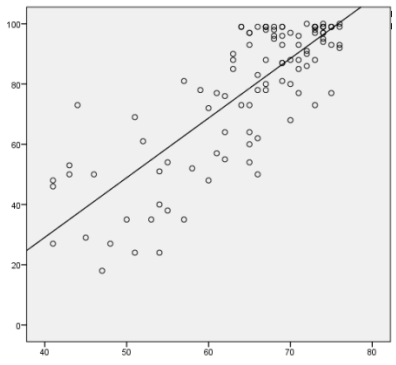
Estas hipóteses estão relacionadas com o significado da regressão. A não rejeição de $H_0 : \beta_1 = 0$ é equivalente à conclusão de que não existe uma relação linear entre $x$ e $Y$. Uma situação possível é ilustrado no Gráfico 3.1. Note-se que isto pode implicar que $x$ é de pouco valor para explicar a variação em $Y$ e que o melhor estimador de $Y$ para qualquer $x$ é $\hat{y} = \hat{Y}$ [Gráfico 3.1 (a)] ou que o verdadeiro relacionamento entre $x$ e $Y$ não é linear [Gráfico 3.2(b)]. Alternativamente, se $H_0 : \beta_1 = 0$ é rejeitado, isto implica que $x$ é de valor para explicar a variabilidade de $Y$ (veja Gráfico 3.2). Rejeitando $H_0 : \beta_1 = 0$ pode significar tanto que o modelo linear é adequado [Gráfico 3.2(a)] ou que, embora haja um efeito linear, os melhores resultados podem ser obtidos com a adição de termos polinomiais de ordem superior em $x$ [Gráfico 3.2(b)].

**Gráfico 3.1** A hipótese $H_0 : \beta_1 = 0$ não é rejeitada

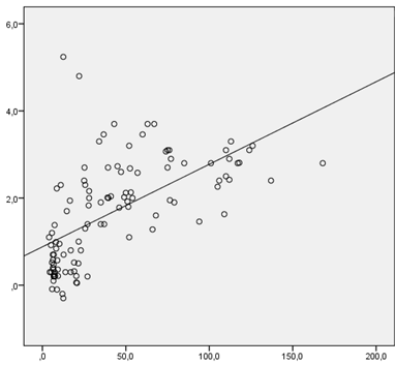



**Fonte:** Própria, dados simulados.

**Gráfico 3.2** A hipótese $H_0 : \beta_1 = 0$ é rejeitada.




(a)




(b)

**Fonte:** Própria, dados simulados.

**Abordagem da Análise de variância para teste de significância de Regressão**

Um método chamado análise de variância (ANOVA) pode ser utilizado para testar a significância de regressão. O procedimento decompõe a variabilidade total da variável resposta em componentes significativas como a base para o ensaio. Desta forma pode-se escrever

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2. \qquad (3.26)$$

As duas componentes do membro direito da equação anterior medem, respectivamente, a quantidade de variabilidade em $y_i$ explicada pela reta de regressão e a variação residual que não é explicada pelo modelo. Usualmente chama-se a $SS_E = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ a soma dos quadrados dos erros e a $SS_R = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ a soma de quadrados da regressão. Simbolicamente, a Equação anterior pode ser reescrita como

$$SS_T = SS_R + SS_E\,, \qquad (3.27)$$

onde $SS_T = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ é a soma dos quadrados total. Note-se que da Equação (3.18) se tem $SS_E = SS_T - \hat{\beta}_1 S_{xy}$, desde que $SS_T = \hat{\beta}_1 S_{xy} + SS_E$. Notemos que a soma de

quadrados da regressão na equação anterior é $SS_R = \hat{\beta}_1 S_{xy}$. A soma total dos quadrados, $SS_T$, tem $n-1$ graus de liberdade, $SS_R$ tem 1 grau de liberdade e $SS_E$ tem $n-2$ graus de liberdade.

Pode-se mostrar que $E\left[SS_E/(n-2)\right] = \sigma^2$, $E(SS_R) = \sigma^2 + \beta_1^2 S_{xx}$ e que $SS_E/\sigma^2$ e $SS_R/\sigma^2$ são variáveis aleatórias independentes com distribuição qui-quadrado com $n-2$ e 1 grau de liberdade, respectivamente. Assim, se a hipótese nula $H_0 : \beta_1 = 0$ se verifica, a estatística

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \tag{3.28}$$

segue uma distribuição $F$ central com 1 e $n-2$ graus de liberdade, $F_{1,n-2}$. Rejeita-se $H_0$ se $f_0 > f_{1-\alpha,1,n-2}$, onde $f_0$ corresponde ao valor observado da estatística e $f_{1-\alpha,1,n-2}$ representa o quartil $1-\alpha$ de distribuição $F$ central com 1 e $n-2$ graus de liberdade. $MS_R = SS_R/1$ e $MS_E = SS_E/(n-2)$ são chamados quadrados médios. Em geral, um quadrado médio é sempre calculado dividindo uma soma de quadrados pelo seu número de graus de liberdade. O procedimento de teste é sempre organizado em uma tabela de análise de variância, tal como a Tabela 3.1.

**Tabela 3.1** Análise da Variância para Testar a Significância da Regressão

| **Fonte de Variação** | **Soma dos Quadrados** | **Graus de Liberdade** | **Quadrados Médios** | $F_0$ |
|---|---|---|---|---|
| Regressão | $SS_R = \hat{\beta}_1 S_{xy}$ | 1 | $MS_R$ | $MS_R / MS_E$ |
| Erro | $SS_E = SS_T - \hat{\beta}_1 S_{xy}$ | $n-2$ | $MS_E$ | |
| Total | $SS_T$ | $n-1$ | | |

**Fonte**: Adaptado de Montegomery & Runger (2011).

Note-se que $MS_E = \hat{\sigma}^2$.

Devemos salientar que o procedimento de análise de variância para testar a significância da regressão linear simples é equivalente ao teste $t$-Student. Ou seja, qualquer dos procedimentos levar-nos-á às mesmas conclusões. Isto é fácil de demonstrar, iniciando com a estatística teste $t$ na Equação (3.21). Com $\beta_{1,0} = 0$, vem

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}. \tag{3.29}$$

Elevando ao quadrado ambos os membros da equação anterior e usando o facto de $\hat{\sigma}^2 = MS_E$, resulta em

$$T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xy}}{MS_E} = \frac{MS_R}{MS_S}. \tag{3.30}$$

Note que $T_0^2$ na Equação (3.30) é idêntico a $F_0$ na Equação (3.28). É verdade, no geral, que o quadrado de uma variável aleatória $t$ com $\nu$ graus de liberdade é uma variável aleatória $F$ com 1 e $\nu$ graus de liberdade. Assim, o teste t é equivalente ao teste $F$ no caso da regressão linear simples.

### 3.2.3 Intervalos de Confiança

**Intervalos de Confiança para os parâmetros do modelo**

Além de apontar estimativas para o declive e intercepto, é possível obter intervalos de confiança para estes parâmetros. A amplitude destes intervalos de confiança é uma medida da qualidade global da reta de regressão. Como visto anteriormente, se os termos do erro, $\varepsilon_i$, no modelo de regressão são normais e independentemente distribuídos, então

$$\left(\hat{\beta}_1 - \beta_1\right) / \sqrt{\hat{\sigma}^2 / S_{xx}} \text{ e } \left(\hat{\beta}_0 - \beta_0\right) / \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

seguem ambos distribuição $t$ com $n-2$ graus de liberdade.

Assim, sob a hipótese de que as observações são normais e independentemente distribuídas, teremos um intervalo de confiança, a um grau de confiança de $100(1-\alpha)\%$ para o declive $\beta_1$ dado por

$$\hat{\beta}_1 - t_{1-\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{1-\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}. \tag{3.31}$$

Similarmente, um intervalo de confiança, a um grau de confiança de $100(1-\alpha)\%$, para $\beta_0$ será dado por

$$\hat{\beta}_0 - t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}. \tag{3.32}$$

**Intervalo de confiança sobre a resposta média**

Um intervalo de confiança pode ser calculado com a resposta média de cada valor de $x$ especificado, digamos, $x_0$. Este intervalo de confiança para $E(Y)=\mu_Y$ é muitas vezes chamado um intervalo de confiança para a reta de regressão. Uma vez que $E(Y)=\mu_Y=\beta_0+\beta_1 x_0$, podemos obter uma estimativa do ponto médio de $Y$ em $x=x_0\,(\mu_Y)$ a partir do modelo ajustado como

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 . \tag{3.33}$$

Agora $\hat{\mu}_Y$ é um estimador não enviesado de $\mu_{Y|x_0}$, uma vez que $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não enviesados de $\beta_0$ e $\beta_1$. A variância de $\hat{\mu}_Y$ será

$$V(\hat{\mu}_Y) = \sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]. \tag{3.34}$$

Este último resultado decorre do facto de que $\hat{\mu}_Y = \bar{y}+\hat{\beta}_1(x_0-\bar{x})$ e $\operatorname{cov}(\bar{Y},\bar{\beta}_1)=0$. Também $\hat{\mu}_{Y|x_0}$ é normalmente distribuído, porque $\hat{\beta}_1$ e $\hat{\beta}_0$ são normalmente distribuídos, e se usarmos $\hat{\sigma}^2$ como um estimador de $\sigma^2$, é fácil mostrar que

$$\frac{\hat{\mu}_Y-\mu_Y}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}},$$

tem distribuição $t$-Student com $n-2$ graus de liberdade. Isto conduz ao seguinte intervalo de confiança.

Um intervalo de confiança, a $100(1-\alpha)\%$, para a resposta média no valor de $x = x_0$, digamos $\mu_Y$, é dado por

$$\hat{\mu}_Y - t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le \mu_Y \le \hat{\mu}_Y + t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}, \qquad (3.35)$$

onde $\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 x_0$ é calculado a partir do modelo de regressão ajustado.

Note-se que a amplitude do intervalo de confiança para $\mu_{Y|x_0}$ é uma função do valor especificado para $x_0$. A amplitude do intervalo é um mínimo para $x_0 = \bar{x}$ e aumenta quando $|x_0 - \bar{x}|$ aumenta.

### 3.2.4 Adequação do modelo de Regressão

A obtenção de um modelo de regressão requer diversos pressupostos. A estimativa dos parâmetros do modelo requer o pressuposto de que os erros são variáveis aleatórias não correlacionadas com média zero e variância constante. Os testes de hipóteses e estimativas de intervalo requerem que os erros sejam normalmente distribuídos. Além disso, assumimos que a ordem do modelo está correta, isto é, se ajustarmos um modelo de regressão linear simples, estamos a supor que o fenómeno realmente se comporta de forma linear.

O investigador deve sempre considerar a validade dessas hipóteses serem duvidosas e realizar análises para examinar a adequação do modelo que foi provisoriamente montado.

Nesta seção, vamos discutir métodos úteis a este respeito.

#### *3.2.4.1 Análise dos Resíduos*

Tanto na Regressão Linear Simples quanto na Regressão Múltipla, os pressupostos do modelo ajustado precisam ser validados para que os resultados sejam confiáveis. Chamamos de **Análise dos Resíduos** a um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. Como visto anteriormente, os erros das observações $\varepsilon_i$, devem ser independentes e

identicamente distribuídos (i.i.d) com distribuição $N\left(0,\sigma^2\right)$. Uma forma de verificar que são validos os pressupostos do modelo consiste em efectuar uma análise dos resíduos, $e_i = y_i - \hat{y}_i$, $i = 1,2,\ldots,n$, onde $y_i$ é um valor observado e $\hat{y}_i$ é um valor ajustado, já que estas são as estimativas dos erros $\varepsilon_i$. Portanto, os resíduos $e_i$ devem refletir as propriedades dos erros: serem normais, com variâncias constantes e independentes.

A seguir apresentamos, dos vários procedimentos possíveis, os mais utilizados para analisar estes pressupostos:

1. Para averiguar se os erros têm distribuição Normal é usual traçar um $PP-plot$ ou um $QQ-plot$ para os resíduos, com base na distribuição normal. O primeiro tipo de gráfico representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros e o segundo representa o quantil de probabilidade esperado se a distribuição fosse normal em função dos resíduos. Também se pode proceder a testes de ajustamento dos resíduos a uma distribuição Normal, nomeadamente os testes de Kolmogorov-Smirnov e Shapiro-Wilk.

    a. Teste de ajustamento de Kolmogorov-Smirnov

    Este teste de ajustamento é dedicado averiguar se determinada amostra provém de uma população contínua com distribuição específica da hipótese nula. A simplicidade deste método advém de ser considerada apenas uma distância, que se pretende tão pequena quanto possível, entre a função de distribuição empririca $F_n(x)$ e a função de distribuição fixada na hipótese nula. Significa então que para $n$ suficientemente grande, as distâncias entre a verdadeira função de distribuição $F$ e o seu representante estatístico $F_n$ dadas por $\left|F_n(x)-F(x)\right|$ deverão ser pequenas para todos os valores de $x$.

    O teste de Kolmogorov-Smirnov tem como estatística de teste

$$D = \sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right| \qquad (3.35)$$

onde $\sup$ designa o supremo e $D$ representa a distância máxima (medida na vertical) entre a função de distribuição empírica e a função de distribuição postulada na hipótese nula.

Um facto de extrema relevância no que diz respeito a esta estatística é o da sua distribuição não depender da distribuição postulada na hipótese nula. Diz-se portanto que a estatística de teste de Kolmogorov-Smirnov é *distribution free*. A função de distribuição aproximada da estatística de Kolmogorov-Smirnov tem uma forma complicada e como tal encontra-se tabelada. Contudo, os pontos críticos permanecem constantes, não dependendo da distribuição a ajustar aos dados (Hall, Neves, & Pereira, 2011).

As hipóteses de interesse, no contexto do teste de ajustamento de Kolmogorov-Smirnov são:

$H_0 : F(x) = F_0(x)$ para todo o $x$ *vs* $H_1 : F(x) \neq F_0(x)$ para algum $x$

Rejeita-se a hipótese de ajustamento desde os valores da estatística de teste sejam elevados. Portanto, a região crítica apropriada, ao nível de significância $\alpha$ é

$$RC_\alpha = \{d : d > c_\alpha\} \tag{3.36}$$

e os valores $c_\alpha$ encontram-se tabelados.

É importante referir que o teste de Kolmogorov-Smirnov quando aplicado a amostras de pequena dimensão tende a não rejeitar a hipótese nula mesmo quando a distribuição que foi postulada não é a que na realidade está subjacente aos dados. Como consequência o teste tende a não rejeitar mesmo que se especifique sucessivamente diferentes distribuições na hipótese nula.

Por outro lado, quando as amotras são grandes (têm dimensão elevada) o teste de Kolmogorov-Smirnov passa a ter comportamento inverso: tende a

rejeitar $H_0$ com grande frequência, sendo quase impossível encontrar evidências de uma distribuição que se ajuste razoalvente aos dados.

Logo, há que ter cuidado na utilização do teste de Kolmogorov-Smirnov e na interpretação dos resultados.

b. Teste de Shapiro-Wilk

Este teste baseia-se numa ideia distinta da do teste de Kolmogorov-Smirnov, isto é, aconselha-se o seu uso para amostras de reduzida dimensão (Hall, Neves, & Pereira, 2011).

2. DIAGNÓSTICO DE INDEPENDÊNCIA

Para verificar se os resíduos são independentes, podemos utilizar técnicas gráficas e testes. A seguir, temos o diagnóstico de independência por meio do teste.

- **Teste de Durbin-Watson**

O teste de Durbin-Watson é utilizado para detetar a presença de auto correlação (dependência) nos resíduos de uma análise de regressão. Este teste é baseado na suposição de que os erros no modelo de regressão são gerados por um processo autorregressivo de primeira ordem, de acordo com

$$\varepsilon_i = \rho\varepsilon_{i-1} + a_i \tag{3.37}$$

em que $\varepsilon_i$ é o termo do erro do modelo na i-ésima observação, $a_i \overset{iid}{\sim} N\left(0, \sigma_a^2\right)$ e $\rho$ $\left(|\rho| < 1\right)$ é o parâmetro de auto correlação. Testamos a presença de auto correlação por meio das hipóteses

$$\begin{cases} H_0 : \rho = 0. \\ H_1 : \rho \neq 0. \end{cases} \tag{3.38}$$

Sendo $e_i$ o resíduo associado à i-ésima observação, temos que a estatística do teste de Durbin-Watson é dada por

$$dw = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (3.39)$$

em que $0 \leq dw \leq 4$. A distribuição de $dw$ depende da matriz $\mathbf{X}$ (matriz do modelo de regressão linear múltipla) . Entretanto, podemos tomar a decisão comparando o valor de $dw$ com os valores críticos $d_L$ e $D_u$ da Tabela de Durbin-Watson (Veja Anexo 2). Assim,

- se $0 \leq dw < d_L$ então rejeitamos $H_0$ (dependência);
- se $d_L \leq dw \leq d_U$ então o teste é inconclusivo;
- se $d_U < dw < 4 - d_U$ então não rejeitamos $H_0$ (independência);
- se $4 - d_U \leq dw \leq 4 - d_L$ então o teste é inconclusivo;
- se $4 - d_L < dw \leq 4$ então rejeitamos $H_0$ (dependência).

Quando $0 \leq dw < d_L$ temos evidência de uma correlação positiva. Já quando $4 - d_L < dw \leq 4$, a correlação é negativa. No caso em que não rejeitamos $H_0$, temos de concluir que não existe autocorrelação, ou seja, os resíduos são independentes. Podemos também tomar a decisão pelo p-valor.

Os valores críticos tabelados apresentados na Tabela de Durbin - Watson (veja o anexo 2) são geralmente utilizados para testar $\rho = 0$ versus $\rho > 0$. Desta forma, se para um determinado $\alpha$ se utilizarem os valores da Tabela de Durbin-Watson (veja o Anexo 5) para testar $\rho = 0$ versus $\rho \neq 0$, o erro tipo I do teste em questão será $2\alpha$.

3. DIAGNÓSTICO DE HOMOSCEDASTICIDADE (VARIÂNCIA CONSTANTE)

Um dos pressupostos do modelo de regressão linear é de que os erros devem ter variância constante. A esse pressuposto chama-se homoscedasticidade.

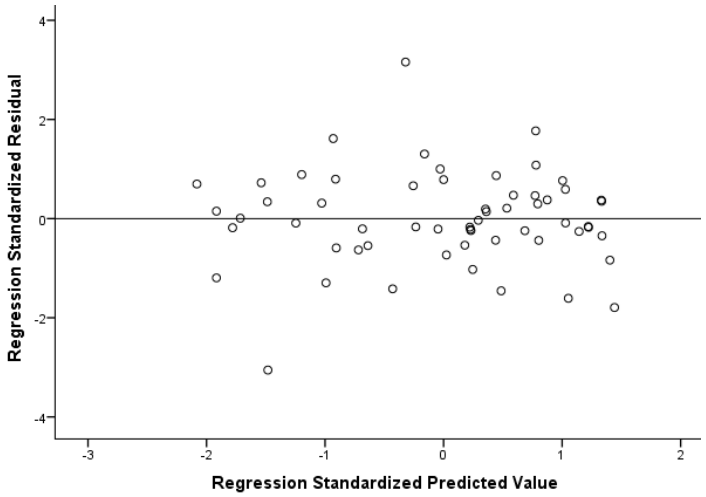
A variância ser constante equivale a supor que não existem observações incluídas na variável residual cuja influência seja mais intensa na variável dependente.

Uma das técnicas utilizadas para verificar a suposição de que os resíduos são homoscedasticos, é a análise do gráfico dos resíduos versus valores ajustados.

a. **Gráfico dos Resíduos versus Valores Ajustados**

Este gráfico deve apresentar pontos dispostos aleatoriamente sem nenhum padrão definido, tal como, por exemplo a Gráfico 3.3.

**Gráfico 3.3** Gráfico de resíduos estandardizados *versus* valores preditos estandardizados




**Fonte:** Própria, dados simulados.

O gráfico dos resíduos versus valores ajustados (valores preditos) é uma das principais técnicas utilizadas para verificar as suposições dos resíduos. Além da detecção de heteroscedasticidade, esse gráfico pode indicar que não existe uma relação linear entra as variáveis explicativas com a variável resposta por meio de alguma tendência nos pontos. Por exemplo, se os pontos do gráfico formam uma parábola, é indicativo que termos de segundo grau sejam necessários.

Para o diagnóstico de heteroscedasticidade, tentamos encontrar alguma tendência no gráfico. Por isso, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento ou tendência, temos indícios de que a variância dos resíduos é homoscedástica. Já a presença de "funil" é um indicativo da presença de heteroscedasticidade.

## 4. DIAGNÓSTICO DE OUTLIERS

*Outlier* é uma observação extrema, ou seja, é um ponto com comportamento diferente dos demais. Além de diagnosticar heteroscedasticidade, o gráfico de resíduos versus valores ajustados também auxilia na detecção de pontos atípicos.

Se um *outlier* for influente, ele interfere sobre a função de regressão ajustada (a inclusão ou não do ponto modifica substancialmente os valores ajustados).

Mas uma observação ser considerada um *outlier* não quer dizer que consequentemente é um ponto influente. Por isso, um ponto pode ser um *outlier* em relação a $Y$ ou aos $X$ , e pode ou não ser um ponto influente.

A detecção de pontos atípicos tem por finalidade identificar:

- *outliers* com relação a $X$ ;
- *outliers* com relação a $Y$ ;
- observações influentes.

### ○ **Outliers em $X$**

*Outliers* em $X$ são detectados por meio da matriz $\mathbf{H} = \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'$ (há ser definida na próxima secção) que transforma o vetor de respostas $Y$ no vetor de valores ajustados $\hat{\mathbf{Y}}$ , pois

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y} \qquad (3.40)$$

Assim, $h_{ii}$ é o i-ésimo elemento da diagonal principal da matriz $\mathbf{H}$ e também é chamado de leverage da observação $i$ .

Além disso, $h_{ii}$ pode ser calculado sem o cálculo completo da matriz H: $h_{ii} = x_i^{'}\left(\mathbf{X}'\mathbf{X}\right)^{-1}x_i$ em que $x_i^{'}$ é a linha da matriz $\mathbf{X}$ correspondente à i-ésima observação.

Podemos identificar observações com alta leverage (*outliers* em $X$ ) das seguintes maneiras:

- Observar se há valores extremos de $h_{ii}$ em um box-plot ou ramo e folhas do mesmo.
- Se a amostra não é grande, considerar $h_{ii}$ grande se é maior que duas vezes a média dos $h_{ii}$. Como $\sum_{i=1}^{n} h_{ii} = posto(\mathbf{X}) = p+1$, a média dos $h_{ii}$ é $(p+1)/n$. É recomendado destacar as observações para as quais $h_{ii} > 2(p+1)/n$.
- Indicar leverage muito alta se $h_{ii} > 0,5$, se a amostra for considerada grande.

- **Outliers em $Y$**

Os resíduos brutos são definidos como $e_i = Y_i - \hat{Y}_i$. Entretanto, para uma melhor detecção em *outliers* em $Y$, eles foram melhorados.

- **Resíduos Padronizados**

O resíduo padronizado, $d_i$, corresponde ao resíduo bruto dividido pelo erro padrão estimado dos resíduos, $\sqrt{QME}$.

$$d_i = \frac{e_i}{\sqrt{QME}} \qquad (3.41)$$

Se os erros têm distribuição normal, então aproximadamente 95% dos resíduos padronizados $d_i$ devem estar no intervalo de [-1.96,1,96,9]. Resíduos fora desse intervalo podem indicar a presença de *outliers*.

- **Resíduos Studentizados**

Existem inúmeras maneiras de se expressar o vetor de resíduos "$\mathbf{e}$" que é útil.

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H}) \qquad (3.42)$$

A matriz de covariâncias dos resíduos é,

$$
\begin{aligned}
Cov[\mathbf{e}] &= Cov\left[(\mathbf{I}-\mathbf{H})\mathbf{Y}\right] = (\mathbf{I}-\mathbf{H})Var(\mathbf{Y})(\mathbf{I}-\mathbf{H})' \\
&= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' \\
&= \sigma^2(\mathbf{I}-\mathbf{H})
\end{aligned} \tag{3.43}
$$

Assim, definimos os resíduos studentizados por

$$
r_i = \frac{e_i}{\sqrt{QME(1-h_{ii})}}, \qquad i = 1,2,\ldots,n, \tag{3.44}
$$

com $\hat{\sigma}^2 = QME$.

Os resíduos studentizados tem variâncias constantes $Var(r_i) = 1$ o que consequentemente torna muito prática a procura por *outliers*, que são observações distantes das demais.

Qualquer observação fora do intervalo $-1,96 \le r_i \le 1,96$ deve ser analisada.

- **Pontos Influentes**

Um ponto é influente se sua exclusão do ajuste da regressão causa uma mudança substancial nos valores ajustados. Por isso, técnicas foram desenvolvidas para identificar essas observações influentes.

- DFFITS

DFFITS mede a influência que a observação $i$ tem sobre seu próprio valor ajustado. Consideremos a medida

$$
DEFITS_{(i)} = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{QME_{(i)}h_{ii}}} \tag{3.45}
$$

isto é, a diferença dos valores preditos de $Y_i$ com e sem a observação $i$ (se $i$ está entre parênteses, significa que é sem essa observação), expressa em unidades de desvios padrões dos valores preditos de $Y_i$.

Assim, essa técnica mede o quanto a inclusão da observação $i$ aumenta ou diminui seu valor predito.

Dizemos que um ponto *outlier* é influente segundo o DFFITS se

- $DEFITS_{(i)} > 1$, para amostras pequenas ou médias.
- $DEFITS_{(i)} > 2\sqrt{p/n}$, para amostras grandes.

➢ DFBETA

DFBETA mede a influência da observação i sobre o coeficiente de $X_j$. É definido por

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QME_i c_{jj}}}, \qquad j = 0,1,\ldots, p. \tag{3.46}$$

em que $c_{jj}$ é o j-ésimo elemento da diagonal de $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

São consideradas observações influentes aquelas que

- $|DFBETA| > 1$, para amostras pequenas.
- $|DFBETA| > 2/\sqrt{n}$, para amostras grandes.

**Distância de Cook**

A distância de Cook mede a influência da observação $i$ sobre todos $n$ valores ajustados $\hat{Y}_i$. É definido por

$$D_i = \frac{e_i^2}{pQME}\frac{h_{ii}}{\left(1-h_{ii}\right)^2}. \tag{3.47}$$

Percebemos que $D_i$ é grande quando ou o resíduo $e_i$ é grande, a leverage $h_{ii}$ é grande ou ambos. Destacamos as observações quando $D_i > 1$.

### 3.2.5 Coeficiente de Determinação

Uma medida amplamente utilizada na regressão é a razão entre a soma dos quadrados. Trata-se do coeficiente de determinação e é dado por

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}. \tag{3.48}$$

Este coeficiente é frequentemente usado para avaliar a adequação de um modelo de regressão aos dados.

O coeficiente de determinação pode ser pensado como uma medida da quantidade de variabilidade explicada (ou tida em conta) pelo modelo de regressão, já que consiste na razão entre a soma dos quadrados dos resíduos e a soma dos quadrados total (Hall, Neves, & Pereira, 2011).

O $R^2$ só pode assumir valores entre 0 e 1. Se existir uma relação linear perfeita entre $Y_i$ e $x_i$ então $R^2 = 1$. Se o declive da recta de regressão for nulo então $R^2 = 0$. Se $R^2 = 0$, então $\beta_1 = 0$, o que significa que não existe relação linear entre as variáveis. Muitas aplicações da análise de regressão envolvem situações em que ambas $X$ e $Y$ são variáveis aleatórias. Nestas situações, é geralmente assumido que as observações $(X_i, Y_i), i = 1, 2, ..., n$ são variáveis aleatórias com distribuição conjunta normal bivariada, e $\mu_Y$ e $\sigma_Y^2$ são a média e a variância de $Y$, $\mu_X$ e $\sigma_X^2$ são a média e a variância de $X$, e $\rho$ é o coeficiente de correlação entre $Y$ e $X$. Lembrando que o coeficiente de correlação é definido por

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{3.49}$$

onde $\sigma_{XY}$ é a covariância entre $Y$ e $X$ e $\sigma_X$ e $\sigma_Y$ correspondem ao desvio padrão de X e Y, respectivamente.

O coeficiente de determinação apresenta um viés positivo em relação ao seu valor na população, o que pode induzir em erro na análise dos dados. Uma forma de compensar este viés consiste em considerar um coeficiente de determinação ajustado definido a partir de $R^2$ e ajustado com base na dimensão da amostra.

$$R^2_{adj} = 1 - \frac{SS_E/(n-2)}{S_{YY}/(n-1)} \tag{3.50}$$

A distribuição condicional de $Y$ para um dado valor de $X = x$ é dada por

$$f_{Y|x}(y) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left[ -\frac{1}{2}\left( \frac{y - \beta_0 - \beta_1 x}{\sigma_{Y|x}} \right)^2 \right] \tag{3.51}$$

onde

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \ ,$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \rho$$

e a variância da distribuição condicionada de $Y$, dado $X = x$, é

$$\sigma^2_{Y|x} = \sigma^2_Y \left(1 - \rho^2\right). \tag{3.52}$$

Isto é, a distribuição condicional de $Y$, dado $X = x$, é normal com média

$$E(Y) = \beta_0 + \beta_1 x \tag{3.53}$$

e variância $\sigma^2_Y$. Assim, o valor esperado da distribuição condicional de $Y$, dado $X = x$, é um modelo de regressão linear simples. Além disso, há uma relação entre o

coeficiente de correlação $\rho$ e o declive $\beta_1$. A partir das equação $\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$ vemos que se $\rho = 0$, então $\beta_1 = 0$, o que implica a ausência de regressão de $Y$ em $X$. Isto é, o conhecimento de $X$ não nos ajuda a prever $Y$.

O método de máxima verossimilhança pode ser usado para estimar os parâmetros $\beta_0$ e $\beta_1$. Pode ser mostrado que os estimadores de máxima verossimilhança dos parâmetros são

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{3.54}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i \left( X_i - \bar{X} \right)}{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2} = \frac{S_{XY}}{S_{XX}} \tag{3.55}$$

Nota-se que os estimadores do intercepto e do declive nas duas equações anteriores são idênticos aos obtidos pelo método dos mínimos quadrados no caso onde $X$ foi assumido como uma variável independente. Isto é, o modelo de regressão com $Y$ e $X$ com distribuição conjunta normal é equivalente ao modelo em que se considera $X$ como uma variável independente. Isto sucede porque as variáveis aleatórias $Y$ dado $X = x$ são independentes e normalmente distribuídas com valor médio $\beta_0 + \beta_1 x$ e variância constante $\sigma^2_{Y|x}$. Estes resultados também serão válidos para qualquer distribuição conjunta de $Y$ e $X$ tal que a distribuição condicional de $Y$ dado $X$ é normal.

É possível realizar inferência sobre o coeficiente de correlação $\rho$ neste modelo. O estimador de $\rho$ é o coeficiente de correlação amostral

$$R = \frac{\sum_{i=1}^{n} Y_i \left( X_i - \bar{X} \right)}{\left[ \sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2 \sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2 \right]^{1/2}} = \frac{S_{XY}}{\left( S_{XX} SS_T \right)^{1/2}} \tag{3.56}$$

Note-se que

$$\hat{\beta}_1 = \left( \frac{SS_T}{S_{XX}} \right)^{1/2} R . \tag{3.57}$$

Assim o declive $\hat{\beta}_1$ é o coeficiente de correlação amostral $R$ multiplicado pelo fator escalar que é a raiz quadrada da "amplitude" dos valores de $Y$ divididos pela amplitude dos valores de $X$. Assim, $\hat{\beta}_1$ e $R$ são intimamente relacionados, embora fornecem informações um pouco diferente. O coeficiente de correlação $R$ mede a associação linear entre $Y$ e $X$, enquanto $\hat{\beta}_1$ mede a mudança predita na média de $Y$ para uma mudança de unidade em $X$. Podemos escrever, a partir da equação anterior que

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{SS_T} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T} \tag{3.58}$$

que é simplesmente o coeficiente de determinação. Isto é, o coeficiente de determinação $R^2$ é o quadrado do coeficiente de correlação entre $Y$ e $X$.

Muitas vezes é útil testar as hipóteses

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned} \tag{3.59}$$

A apropriada estatística de teste para estas hipóteses será

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}, \tag{3.60}$$

que segue uma distribuição $t$-Student com $n-2$ graus de liberdade, sob $H_0 : \rho = 0$. Portanto, rejeitaríamos a hipótese nula se $|t_0| > t_{1-\alpha/2,n-2}$. Este teste é equivalente ao teste de hipótese para $\beta_1$, com $H_0: \beta_1 = 0$, visto anteriormente. Esta equivalência resulta directamente da equação anterior.

O procedimento do teste para as hipóteses

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0 \end{aligned} \tag{3.61}$$

onde $\rho_0 \neq 0$ é um pouco mais complicado. Para amostras moderadamente grandes (digamos, $n \geq 25$), a estatística (Montegomery & Runger, 2011, p. 436),

$$Z = \text{arct anh R} = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \tag{3.62}$$

segue uma distribuição aproximadamente normal com média e variância

$$\mu_Z = \text{arctan}\, h\ \rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \text{ e } \sigma_Z^2 = \frac{1}{n-3}\ ,$$

respetivamente. Portanto, para testar a hipótese $H_0 : \rho = \rho_0$, podemos usar a estatística de teste

$$Z_0 = \left(\text{arctan h}\ R - \text{arctan h}\ \rho_0\right)\left(n-3\right)^{1/2}. \tag{3.63}$$

Rejeita-se $H_0 : \rho = \rho_0$ se o valor da estatística de teste na equação anterior é tal que $|z_0| > z_{1-\alpha/2}$, em que $z_{1-\alpha/2}$ corresponde ao quantil $1-\alpha/2$ da distribuição da estatística. É também possível construir um intervalo de confiança para $\rho$, usando

transformações na equação $Z = arc\,\text{tagh}\ \text{R} = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$. Assim, o intervalo de confiança, a um grau de confiança de $100(1-\alpha)\%$, será então

$$\tanh\left(\text{arctan h}\ r - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\text{arctan h}\ r + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right), \tag{3.64}$$

onde $\tanh u = \left(e^{u} - e^{-u}\right) / \left(e^{u} + e^{-u}\right)$ e $\text{arctanh}$ é a inversa da tangente hiperbólica.

## 3.3 Regressão Linear Múltipla

### 3.3.1 Introdução

Muitas aplicações da análise de regressão envolvem situações em que existem mais do que uma variável regressora ou preditor. Um modelo de regressão que contenha mais que uma variável regressora recebe o nome de modelo de regressão múltipla.

Suponhamos, por exemplo, que a vida útil de uma ferramenta de corte depende da velocidade de corte e do ângulo da ferramenta. Um modelo de regressão múltipla que poderia descrever esta relação é

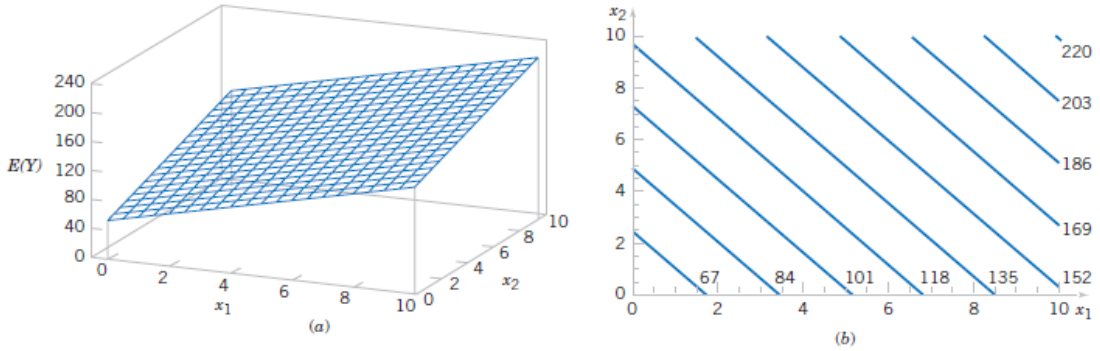
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{3.65}$$

onde $Y$ representa a vida da ferramenta, $x_1$ representa a velocidade de corte, $x_2$ representa o ângulo da ferramenta e $\varepsilon$ é o termo aleatório do erro. Este é um modelo de regressão linear múltiplo com duas regressoras. O termo linear é usado porque trata-se de uma função linear com parâmetros desconhecidos $\beta_0, \beta_1$ e $\beta_2$.

O modelo de regressão na equação anterior descreve um plano em três dimensões no espaço de $Y$, $x_1$ e $x_2$. A Figura 14 mostra este plano para o modelo

$$E(Y) = 50 + 10x_1 + 7x_2,$$

**Figura 3.1** (a) Plano de regressão para o modelo $E(Y) = 50 + 10x_1 + 7x_2$. (b) diagrama de contorno.




**Fonte:** Adaptado de Montegomery & Runger (2011)

onde assumimos que o valor esperado do termo do erro é zero; isto é $E(\varepsilon) = 0$. O parâmetro $\beta_0$ é o intercepo do plano. Por vezes chamamos a $\beta_1$ e $\beta_2$ coeficientes parciais de regressão, porque, $\beta_1$ mede a mudança esperada em $Y$ por unidade de mudança em $x_1$ quando $x_2$ é mantido constante, e $\beta_2$ mede a mudança em $Y$ por unidade em $x_2$ quando $x_1$ é mantido constante. A Figura 3.1(b) mostra o diagrama de contorno do modelo de regressão, isto é, linhas constantes $E(Y)$ com uma função de $x_1$ e $x_2$. Note-se que as linhas de contorno neste diagrama são linhas rectas.

Em geral, a variavel dependente ou resposta $Y$ pode estar relacionada a $k$ variáveis independentes ou regressoras. O modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3.66}$$

é chamado modelo de regressão linear múltipla com $k$ variáveis regressoras. Os parâmetros $\beta_j$, $j = 0,1,\ldots,k$, são chamados coeficientes de regressão. Este modelo descreve um hiperplano da mudança esperada na resposta $Y$ por unidade de mudança em $x_j$ quando todas as demais regressoras são mantidas constantes.

Os modelos de regressão linear são frequetemente usados como funções aproximadas, isto é, a verdadeira relação funcional entre $Y$ e $x_1, x_2, \ldots, x_k$ é desconhecida, mas ao longo de determinados intervalos das variáveis independentes o modelo de regressão linear é uma aproximação adequada.

Modelos mais complexos do que o apresentado anteriormente podem frequentemente ser analisados por várias técnicas de regressão linear. Por exemplo, consideremos o modelo polinomial cúbico com uma única variável regressora.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

Se fizermos $x_1 = x, x_2 = x^2, x_3 = x^3$, teremos:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon ,$$

que é um modelo de regressão linear múltipla com três variáveis regressoras.

Os modelos que incluam efeitos de interação podem também ser analizados por métodos da regressão linear múltipla. Uma interação entre duas variáveis pode ser representada por um termo de produto cruzado no modelo, tal como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

Se fizermos $x_3 = x_1 x_2$ e $\beta_3 = \beta_{12}$, a equação anterior pode ser escrita como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

que é um modelo de regressão linear.

As Figuras 3.2(a) e (b) mostram um gráfico tridimensional do modelo de regressão

$$Y = 50 + 10 x_1 + 7 x_2 + 5 x_1 x_2$$

e o correspondente diagrama de contorno. Nota-se que, embora este modelo seja um modelo linear, a forma da superfície que é gerada pelo modelo não é linear. Em geral, qualquer modelo que é linear nos parâmetros ($\beta$'s) é um modelo de regressão linear, independentemente da forma da superficie que gera.

A Figura 3.2 fornece uma boa interpretação gráfica de uma interação. Geralmente, interações implicam que os efeitos produzidos pela mudança de uma variável ($x_1$, digamos) depende do nível da outra variável ($x_2$). Por exemplo, a Figura 3.2 mostra que mudando $x_1$ de 2 para 8 produz uma mudança mais pequena em $E(Y)$ quando $x_2 = 2$ do que quando $x_2 = 10$. Os efeitos de interação ocorrem frequentemente em estudos e análises de sistema do mundo real, e os métodos de regressão são uma das técnicas que podemos usar para descrevê-los.

Como um exemplo final, consideremos o modelo de segunda ordem com interação

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

Se fizermos $x_3 = x_1^2, x_4 = x_2^2, x_5 = x_1 x_2, \beta_3 = \beta_{11}, \beta_4 = \beta_{22}$ e $\beta_5 = \beta_{12}$, a equação anterior pode ser escrita como um modelo de regressão linear múltipla como segue:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

A Figura 3.3(a) e (b) mostra um gráfico tridimensional e o correspondente diagrama de contorno para

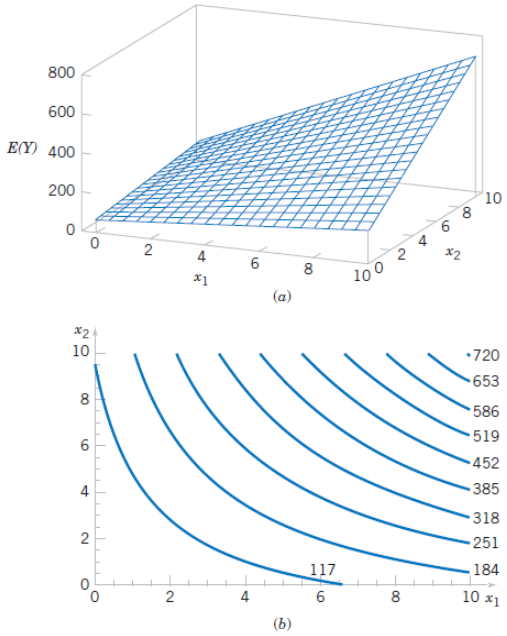
$$E(Y) = 800 + 10x_1 + 10x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1 x_2$$

Estes gráficos indicam que as mudanças esperadas em $Y$ quando $x_i$ é mudado por uma unidade, é uma função de ambos $x_1$ e $x_2$. Os termos quadráticos e de interação no modelo produzem uma função em forma de colina.

### 3.3.2 Estimativas dos Parâmetros dos Mínimos Quadrados

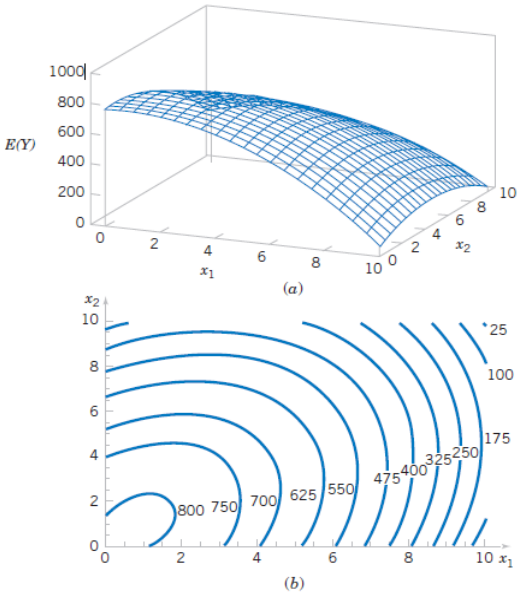
O método dos mínimos quadrados pode ser usado para estimar os coeficientes de regressão no modelo de regressão múltipla. Suponha que $n > k$ observações estão disponíveis, e seja $x_{ij}$ a i-ésima observação ou nível da variável $x_j$. As observações são

**Figura 3.2** (a) Gráfico Tridimensional do modelo de regressão $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. (b) Diagrama de contornos.




**Fonte:** Adaptado de Montegomery & Runger (2011)

**Figura 3.3** (a) ) Gráfico Tridimensional do modelo de regressão $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) Diagrama de contornos.




**Fonte:** Adaptado de Montegomery & Runger (2011)

$$(x_{i1}, x_{i2}, \ldots x_{ik}, y_i), \; i = 1, 2, \ldots, n \text{ e } n > k.$$

É costume apresentar os dados para a regressão múltipla em uma tabela tal como a Tabela 3.2

Cada observação $(x_{i1}, x_{i2}, \ldots x_{ik}, y_i)$, satisfaz o modelo na equação do modelo de regressão linear múltipla, ou

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad , i = 1, 2, \ldots, n \end{aligned} \tag{3.67}$$

**Tabela 3.2** Dados para a Regressão Linear Múltipla

| $y$ | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| $y_1$ | $x_{11}$ | $x_{12}$ | … | $x_{1k}$ |
| $y_2$ | $x_{21}$ | $x_{22}$ | … | $x_{2k}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ |
| $y_n$ | $x_{n1}$ | $x_{n2}$ | … | $x_{nk}$ |

A função dos mínimos quadrados é

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 . \tag{3.68}$$

Desejamos minimizar $L$ com respeito a $\beta_0, \beta_1, \ldots, \beta_k$. As estimativas dos mínimos quadrados de $\beta_0 . \beta_1, \ldots, \beta_k$ devem satisfazer

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_k} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{3.69a}$$

e

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_k} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 , \quad j = 1, 2, \ldots, k \tag{3.69b}$$

Simplificando a Equação (3.69), obtemos as equações normais dos mínimos quadrados.

$$
\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
\tag{3.70}
$$

Nota-se que há $p = k + 1$ equações normais, uma para cada um dos coeficientes desconhecidos de regressão. A solução para as equações normais serão os estimadores dos mínimos quadrados dos coeficientes de regressão, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. As equações normais podem ser resolvidas por métodos apropriados para a resolução de sistemas de equações lineares.

### 3.3.3 Abordagem Matricial para um Modelo de Regressão Linear Múltipla

No ajustamento de um modelo de regressão múltipla, é muito mais conveniente expressar as operações matemáticas usando notações matriciais. Suponhamos que existem $k$ variáveis regressoras e $n$ observações, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), i = 1, 2, \ldots, n$ e que o modelo que relaciona as regressoras e a resposta é

$$
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad , i = 1, 2, \ldots, n \tag{3.71}
$$

Este modelo é um sistema de $n$ equações que pode ser expresso em notação matricial como

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \, , \tag{3.72}
$$

onde

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

Em geral, $\mathbf{y}$ é um vetor $(n\times 1)$ das observações, $\mathbf{X}$ é uma matriz $(n\times p)$ dos níveis das variáveis independentes (assuma-se que o intercepto é sempre multiplicado por um valor constante – unidade), $\boldsymbol{\beta}$ é um vetor $(p\times 1)$ dos coeficientes de regressão, $p=k+1$ e $\boldsymbol{\varepsilon}$ é um vetor $(n\times 1)$ dos erros aleatórios. A matriz $\mathbf{X}$ é geralmente chamada matriz do modelo.

Queremos encontrar o vetor dos estimadores dos mínimos quadrados, $\hat{\beta}$, que minimiza

$$L=\sum_{i=1}^{n}\varepsilon_i^2=\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{3.73}$$

O estimador dos mínimos quadrados $\hat{\boldsymbol{\beta}}$ é a solução para $\boldsymbol{\beta}$ nas equações

$$\frac{\partial L}{\partial \boldsymbol{\beta}}=0 \tag{3.74}$$

Mesmo sem demonstração, as equações resultantes que devem ser resolvidas são

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y} \tag{3.75}$$

As equações anteriores são as equações normais dos mínimos quadrados em forma de matriz. Para resolver as equações normais, multiplicam-se ambos os membros destas equações pelo inverso de $\mathbf{X}'\mathbf{X}$. Então, a estimativa dos mínimos quadrados de $\boldsymbol{\beta}$ é

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.76}$$

Note-se que há $p = k + 1$ equações normais e $p = k + 1$ parâmetros desconhecidos (os valores de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$). Além disso, a matriz $\mathbf{X}'\mathbf{X}$ é sempre não singular, como foi assumido acima.

É fácil verificar que a forma matricial das equações normais é idêntica à forma escalar. Escrevendo a fórmula das equações normais dos mínimos quadrados em detalhes, obtém-se

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}. \qquad (3.77)$$

O modelo ajustado de regressão é

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}\ , i = 1, 2, \ldots, n. \qquad (3.78)$$

Em notação matricial, do modelo ajustado é

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \qquad (3.79)$$

A diferença entre as observações e o valor ajustado $\hat{y}_i$ é um resíduo, isto é, $e_i = y_i - \hat{y}_i$. O vetor $(n \times 1)$ dos resíduos é denotado por

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \qquad (3.80)$$

### 3.3.4 Propriedades dos Estimadores dos Mínimos Quadrados

As propriedades estatísticas dos estimadores dos mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ podem ser facilmente encontradas, sob certas hipóteses, sobre os termos dos erros $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, no modelo de regressão. Paralelamente as hipóteses apresentadas aquando da abordagem da regressão linear simples, assumimos que os erros $\varepsilon_i$ são estatisticamente independentes com média zero e variância $\sigma^2$. Sob estes pressupostos, os estimadores dos mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ são estimadores não enviesados dos coeficientes de regressão $\beta_0, \beta_1, \ldots, \beta_k$. Estas propriedades podem ser apresentadas como se segue:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}\right) &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta}+\varepsilon\right)\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\right] \\ &= \boldsymbol{\beta} \end{aligned} \tag{3.81}$$

uma vez que $E(\varepsilon)=0$ e $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}=\mathbf{I}$, onde $\mathbf{I}$ representa a matriz identidade. Assim, $\hat{\boldsymbol{\beta}}$ é um estimador centrado de $\boldsymbol{\beta}$.As variâncias dos $\hat{\boldsymbol{\beta}}$'s são expressas em termos dos elementos da inversa da matriz $\mathbf{X}'\mathbf{X}$. A inversa de $\mathbf{X}'\mathbf{X}$ vezes a constante $\sigma^2$ representa a matriz de covariância dos coeficientes de regressão. Os elementos da diagonal de $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ são as variâncias de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ e os elementos fora desta diagonal são as covariâncias.

Em geral, a matriz de covariância de $\hat{\boldsymbol{\beta}}$ é uma matriz simétrica $(p \times p)$ da qual o $jj$-ésimo elemento é a variância de $\hat{\beta}_j$ e da qual o $i,j$-ésimo elemento é a covariância entre $\hat{\beta}_i$ e $\hat{\beta}_j$, isto é,

$$\text{cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \mathbf{C}. \qquad (3.82)$$

onde $\mathbf{C} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

As estimativas das variâncias dos coeficientes de regressão são obtidas pela substituição de $\sigma^2$ por um estimador Quando $\sigma^2$ é substituído pelo seu estimador $\hat{\sigma}^2$, a raiz quadrada das variâncias estimadas do $j$-ésimo coeficiente de regressão é chamado de erro padrão estimado de $\hat{\beta}_j$ ou $se\left(\hat{\beta}_j\right) = \sqrt{\hat{\sigma}^2 C_{jj}}$. Estes erros padrões são uma medida útil da precisão da estimação para os coeficientes de regressão: Erros padrões pequenos ou baixos implicam uma boa precisão.

### 3.3.5 Estimador de $\sigma^2$

Tal como na regressão linear simples, é importante estimar $\sigma^2$, a variância do termo de erro, $\varepsilon$, num modelo de regressão linear múltipla. Recordemos que na regressão linear simples o estimador de $\sigma^2$ foi obtido dividindo a soma do quadrado dos resíduos por $n-2$. Assim considerando uma regressão linear múltipla com $p$ parâmetros um estimador lógico para $\sigma^2$ será

$$\hat{\sigma}^2 = \frac{\sum_{i=1} e_i^2}{n-p} = \frac{SS_E}{n-p}. \qquad (3.83)$$

Este é um estimador não enviesado (centrado) de $\sigma^2$. Tal como em uma regressão linear simples, o estimador de $\sigma^2$ é habitual ser obtido a partir da análise de variância para o modelo de regressão linear. O numerador da equação anterior corresponde à soma dos quadrados dos resíduos e o denominador, $n-p$, aos graus de liberdade do resíduo.

Podemos encontrar uma fórmula de cálculo para $SS_E$ como se segue:

$$SS_E = \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}. \tag{3.84}$$

Substituindo $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ na equação acima, obtemos

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{3.85}$$

### 3.3.6 Testes de Hipóteses em Regressão Linear Múltipla

Em problemas de regressão linear múltipla, certos testes de hipóteses acerca dos parâmetros do modelo são uteis na medição da adequação do modelo. Nesta secção descreveremos vários procedimentos de testes de hipóteses importantes. Como no caso da regressão linear simples, requere-se que os termos dos erros no modelo de regressão sejam independentes e normalmente distribuídos com média zero e variância $\sigma^2$.

#### *3.3.6.1 Testes para a Significância da Regressão*

O teste para significância da regressão é um teste para determinar se existe uma relação linear entre a variável resposta $y$ e um conjunto de variáveis regressoras $x_1, x_2, \ldots, x_k$. As hipóteses apropriadas são

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1 : \beta_j \neq 0 \text{ para pelo menos um } j \end{aligned} \tag{3.86}$$

A rejeição da $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implica que pelo menos uma das variáveis regressoras $x_1, x_2, \ldots, x_k$ é significativa, constribuindo para a explicação da variável dependente.

O teste de significância da regressão é uma generalização dos procedimentos usados na regressão linear simples. A soma dos quadrados total $SS_T$ é particionada em uma soma de quadrados devido ao modelo ou à regressão e uma soma de quadrados devido ao erro, digamos,

$$SS_T = SS_R + SS_E . \tag{3.87}$$

Agora se $H_0$ é verdadeira, $SS_R / \sigma^2$ é uma variável aleatória com distribuição qui-quadrado com $k$ graus de liberdade. Note-se que o número de graus de liberdade para esta variável aleatória é igual ao número de variáveis regressoras no modelo. Podemos também mostrar que $SS_E / \sigma^2$ é um qui-quadrado com $n-k-1$ graus de liberdade, e que $SS_E$ e $SS_R$ são independentes. A estatística teste para $H_0$ é

$$F_0 = \frac{SS_R / k}{SS_E / (n-k-1)} = \frac{MS_R}{MS_E} . \tag{3.88}$$

Devemos rejeitar $H_0$ se o valor observado da estatística $F_0$ , $f_0$, for maior do que $f_{1-\alpha,k,n-k-1}$. O procedimento é geralmente resumido na tabela da ANOVA, tal como a Tabela 3.3

Uma fórmula de cálculo para $SS_R$ pode ser encontrada facilmente. Agora, uma vez que $SS_T = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n = \mathbf{y}'\mathbf{y} - \left(\sum_{i=1}^{n} y_i\right)^2 / n$, podemos reescrever a Equação (3.88) como

$$SS_E = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \tag{3.89}$$

**Tabela 3.3** Análise da Variância para Testar a Significância da Regressão em Análise de Regressão Múltipla

| **Fonte de Variação** | **Soma dos Quadrados** | **Graus de Liberdade** | **Quadrados Médios** | $F_0$ |
|---|---|---|---|---|
| Regressão | $SS_R$ | $k$ | $MS_R$ | $MS_R$ / $MS_E$ |
| Erro ou resíduo | $SS_E$ | $n-k-1$ | $MS_E$ | |
| Total | $SS_T$ | $n-1$ | | |

**Fonte:** Adaptado de Montegomery & Runger (2011).

ou

$$SS_E = SS_T - SS_R . \tag{3.90}$$

A soma dos quadrados da regressão é

$$SS_R = \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} . \tag{3.91}$$

### *3.3.6.2 O $R^2$ e o $R^2$ ajustado*

Também se pode usar o coeficiente de determinação múltiplo $R^2$ como uma estatística global para avaliar o ajuste do modelo. Calcula-se,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} . \tag{3.92}$$

Muitos utilizadores da regressão preferem usar uma estatística ajustada do $R^2$:

$$R^2_{\text{adj}} = 1 - \frac{SS_E / (n-p)}{SS_T / (n-1)}. \tag{3.93}$$

Como $SS_E / (n-p)$ corresponde ao quadrado médio residual e $SS_T / (n-1)$ é uma constante, $R^2_{\text{adj}}$ apenas aumenta quando a variável que é adicionada ao modelo reduz o quadrado da média do erro.

A estatística ajustada de $R^2$ essencialmente penaliza o investigador por adicionar termos ao modelo. É uma maneira fácil de se proteger contra superajuste, isto é, incluir variáveis regressoras que não são realmente úteis. Consequentemente, é muito útil para a avaliação do modelo de regressão.

#### *3.3.6.3 Testes em Coeficientes Individuais de Regressão e Subconjuntos de Coeficientes*

Estamos frequentemente interessados em testar hipóteses sobre os coeficientes individuais da regressão. Tais testes seriam úteis para determinar o potencial valor de cada variável regressora do modelo de regressão. Por exemplo, o modelo pode ser mais efetivo com a inclusão de variáveis adicionais ou talvez com a eliminação de uma ou mais regressoras presentes no modelo.

As hipóteses para testar que um coeficiente individual de regressão, digamos $\beta_j$, *j=1,...,k,* é igual a um valor $\beta_{j0}$, são

$$\begin{aligned} H_0 &: \beta_j = \beta_{j0} \\ H_1 &: \beta_j \neq \beta_{j0} \end{aligned}. \tag{3.94}$$

A estatística teste para esta hipótese é

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{se\left(\hat{\beta}_j\right)}, \tag{3.95}$$

onde $C_{jj}$ é o elemento da diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$ correspondente a $\hat{\beta}_j$. Note-se que o denominador da equação anterior é o erro padrão do coeficiente de regressão $\hat{\beta}_j$. A hipótese nula $H_0 : \beta_j = \beta_{j0}$ é rejeitada se $|t_0| > t_{1-\alpha/2,n-p}$. Este é chamado um teste parcial ou marginal porque o coeficiente de regressão $\hat{\beta}_j$, *j=1,...,k*, depende de todas as outras variáveis regressoras $x_i \, (i \neq j)$ que estão no modelo.

Um caso particular importante das hipóteses anteriores ocorre para $\beta_{j0} = 0$. Se $\beta_j = 0$ não é rejeitada, isto indica que a regressora $x_j$ pode ser eliminada do modelo. Acrescentando uma variável ao modelo de regressão faz sempre com que a soma dos quadrados da regressão aumente e a soma dos quadrados do erro diminua (esta é a razão que faz com que $R^2$ aumente sempre quando se acrescenta uma variável ao modelo). Devemos decidir se o aumento da soma dos quadrados da regressão é grande o suficiente para justificar o uso de variáveis adicionais no modelo. Além disso, adicionando uma variável não importante ao modelo pode realmente aumentar o quadrado da média do erro, indicando que a adição de tal variável realmente torna o modelo pobre em termos de ajuste aos dados (eis a razão de $R^2_{\text{adj}}$ ser a melhor medida do ajustamento global de um modelo de regressão linear do que o habitual $R^2$).

Existe uma outra maneira para testar a contribuição das variáveis individuais de regressão. Esta abordagem determina o aumento na soma dos quadrados da regressão obtido, adicionando uma variável $x_j$, $j = 1,\ldots,k$, ao modelo, dado que outras variáveis $x_i \, (i \neq j)$ estão já incluídas na equação de regressão.

O procedimento usado para fazer isto é chamado teste geral de significância da regressão, ou método extra da soma dos quadrados. Este procedimento pode também ser usado para investigar a contribuição do subconjunto das variáveis regressoras do modelo. Consideremos o modelo de regressão com $k$ variáveis regressoras

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad (3.96)$$

onde $\mathbf{y}$ é $(n\times 1)$, $\mathbf{X}$ é $(n\times p)$, $\boldsymbol{\beta}$ é $(p\times 1)$, $\varepsilon$ é $(n\times 1)$ e $p=k+1$. Gostaríamos de determinar se os subconjuntos de variáveis regressoras $x_1, x_2, \ldots x_r\ (r<k)$ como um todo, contribuiem significativamente para o modelo de regressão. Seja o vetor dos coeficientes de regressão particionados como se segue:

$$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}, \tag{3.97}$$

onde $\boldsymbol{\beta}_1$ é $(r\times 1)$ e $\boldsymbol{\beta}_2$ é $\left[(p-r)\times 1\right]$. Queremos testar as hipóteses

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_1=\mathbf{0}\\ H_1 &: \boldsymbol{\beta}_2\neq\mathbf{0}\end{aligned}, \tag{3.98}$$

onde $\mathbf{0}$ denota o vetor nulo. O modelo pode ser escrito como

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\grave{\mathbf{o}}\equiv\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}. \tag{3.99}$$

onde $\mathbf{X}_1$ representa as colunas de $\mathbf{X}$ associado a $\boldsymbol{\beta}_1$ e $\mathbf{X}_2$ representa as colunas de $\mathbf{X}$ associado a $\boldsymbol{\beta}_2$.

Para um modelo completo (incluindo $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$), sabemos que $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. Além disso, a soma de quadrados da regressão para todas as variáveis, incluindo o intercepto, são

$$SS_R\left(\boldsymbol{\beta}\right)=\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}\quad\left(p=k+1\text{ graus de liberdade}\right) \tag{3.100}$$

e

$$MS_E = \frac{\mathbf{y'y - \hat{\beta}X'y}}{n-p}. \tag{3.101}$$

$SS_R(\boldsymbol{\beta})$ é chamado a soma dos quadrados da regressão devido a $\boldsymbol{\beta}$. Para encontrar o contributo dos termos em $\boldsymbol{\beta}_1$ à regressão, ajustar o modelo assumindo a hipótese $H_0 : \boldsymbol{\beta}_1 = 0$ seja verdadeira. O modelo reduzido é encontrado a partir da Equação (3.99) como

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{3.102}$$

O estimador dos mínimos quadrados para $\boldsymbol{\beta}_2$ é $\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X'y}$, e

$$SS_R\left(\boldsymbol{\beta}_2\right) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} \quad \left(p-r \text{ graus de liberdade}\right). \tag{3.103}$$

A soma dos quadrados da regressão devido a $\boldsymbol{\beta}_1$, dado que $\boldsymbol{\beta}_2$ já está no modelo, é

$$SS_R\left(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2\right) = SS_R\left(\boldsymbol{\beta}_1\right) - SS_R\left(\boldsymbol{\beta}_2\right) \tag{3.104}$$

Esta soma dos quadrados tem $r$ graus de liberdade. É algumas vezes chamada de soma dos quadrados extra devido a $\boldsymbol{\beta}_1$. Nota-se que $SS_R\left(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2\right)$ é o incremento da soma dos quadrados da regressão devido às variáveis $x_1, x_2, \ldots, x_r$ no modelo. Agora $SS_R\left(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2\right)$ é independente de $MS_E$, e a hipótese nula $\boldsymbol{\beta}_1 = \mathbf{0}$ pode ser testada pela estatística

$$F_0 = \frac{SS_R\left(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2\right)/r}{MS_E} \tag{3.105}$$

Se o valor calculado da estatística teste $f_0 > f_{1-\alpha,r,n-p}$, rejeitamos $H_0$, concluindo que pelo menos um dos parâmetros em $\boldsymbol{\beta}_1$ não é zero e, consequentemente, pelo menos uma

das variáveis $x_1, x_2, \ldots, x_r$ em $\mathbf{X_1}$ contribui significativamente para o modelo de regressão. A fórmula anterior é chamada por alguns autores como sendo o teste $F$ parcial.

O teste $F$ parcial é muito útil. Podemos, usá-lo para medir a contribuição individual de cada uma das regressoras $x_j$, $j = 1, \ldots, k$ como se fosse a última variável adicionada ao modelo pelo cálculo,

$$SS_R\left(\beta_j \mid \beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right), \qquad j = 1.2. \ldots, k \ .$$

Este é o incremento na soma dos quadrados da regressão devido à adição de $x_j$, $j = 1, \ldots, k$, ao modelo que já inclui $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$. O teste $F$ parcial é um procedimento muito genérico em que podemos medir o efeito dos conjuntos de variáveis.

### 3.3.7 Intervalos de Confiança na Regressão Linear Múltipla

#### *3.3.7.1 Intervalos de Confiança para os Coeficientes Individuais de Regressão*

Em modelos de regressão múltipla, é útil construir intervalo de confiança (IC) para os coeficientes de regressão, $\left(\beta_j,\ j = 1, \ldots, k\right)$ . No desenvolvimento de um procedimento para obter estes intervalos de confiança é necessário que os erros $\varepsilon_i$ sejam independentes e normalmente distribuídos com média zero e variância $\sigma^2$. Este é o pressuposto necessário no teste de hipótese. Portanto, as observações $\{Y_i\}$ são independentes e normalmente distribuídas com média $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ e variância $\sigma^2$. Uma vez que o estimador dos mínimos quadrados $\hat{\boldsymbol{\beta}}$ é uma combinação linear das observações, segue-se que $\hat{\boldsymbol{\beta}}$ é normalmente distribuído com o vetor médio $\boldsymbol{\beta}$ e matriz de covariância $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$. Então cada uma das estatísticas

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad j = 0, 1, \ldots, k \tag{3.106}$$

tem uma distribuição $t$ com $n-p$ graus de liberdade, onde $C_{jj}$ é o $j$-ésimo elemento de $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, e $\hat{\sigma}^2$ é o estimador da variância do erro, obtido a partir da Equação (3.83).

Assim, um intervalo de confiança, a um grau de confiança de $100(1-\alpha)\%$, para os coeficientes de regressão $\beta_j$, $j=0,1,\ldots,k$, no modelo de regressão linear é dado por

$$\hat{\beta}_j - t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{3.107}$$

.

Como $\sqrt{\hat{\sigma}^2 C_{jj}}$ é o erro padrão do coeficiente de regressão $\hat{\beta}_j$, poderíamos também escrever a fórmula do intervalo de confiança como

$$\hat{\beta}_j - t_{1-\alpha/2,n-p}se\left(\hat{\beta}_j\right) \le \beta_j \le \hat{\beta}_j + t_{1-\alpha/2,n-p}se\left(\hat{\beta}_j\right).$$

#### *3.3.7.2 Intervalo de Confiança para a Resposta Média*

Podemos também obter um intervalo de confiança para a resposta média em um ponto particular, digamos, $x_{01}, x_{02}, \ldots, x_{0k}$. Para estimar a resposta média neste ponto, definimos o vetor

$$\mathbf{x_0} = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}. \tag{3.108}$$

A resposta média neste ponto é $E\left(Y \mid \mathbf{X}_0\right) = \mu_{Y|\mathbf{X}_0} = \mathbf{X}_0\boldsymbol{\beta}$, que é estimada por

$$\hat{\boldsymbol{\mu}}_{Y|\mathbf{x_0}} = \mathbf{x_0}'\hat{\boldsymbol{\beta}}. \tag{3.109}$$

O estimador é não enviesado uma vez que $E\left(\mathbf{x}_0'\hat{\boldsymbol{\beta}}\right) = \mathbf{x}_0'\boldsymbol{\beta} = E\left(Y \mid \mathbf{x_0}\right) = \mu_{Y|\mathbf{x}_0}$ e a variância de $\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}_0}$ é

$$V\left(\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}_0}\right) = \sigma^2\mathbf{x}_0'\left(\mathbf{X'X}\right)^{-1}\mathbf{x}_0 . \tag{3.110}$$

Um IC, a um grau de confiança $100(1-\alpha)\%$, para $\mu_{Y|\mathbf{x}_0}$ pode ser construído a partir da estatística

$$\frac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2\mathbf{x}_0'\left(\mathbf{X'X}\right)^{-1}\mathbf{x}_0}} . \tag{3.111}$$

Assim, um intervalo de confiança, para um grau de confiança de $100(1-\alpha)\%$, na resposta média no ponto $x_{01}, x_{02}, \ldots, x_{0k}$ é

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'\left(\mathbf{X'X}\right)^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'\left(\mathbf{X'X}\right)^{-1}\mathbf{x}_0} \tag{3.112}$$

A Equação (3.112) é um intervalo de confiança do plano de regressão (híper-plano).

### 3.3.8 Avaliação da qualidade e significado da regressão

Para avaliar a qualidade e significado da regressão deve-se novamente considerar vários métodos.

- Métodos gráficos

Uma vez que o modelo de regressão múltipla tem por base a equação de uma superfície, não é fácil visualizar graficamente as observações $Y$ em função das várias variáveis regressoras individualmente. É também habitual construir um gráfico de dispersão que apresente os valores observados $Y_i$ versus os valores preditos $\hat{Y}_i$. Tal como no modelo de regressão simples, se o modelo for adequado, os valores preditos devem estar próximos dos valores observados e portanto este gráfico de dispersão deve apresentar um conjunto de pontos próximos da recta $y = x$.

- Coeficiente de determinação

O coeficiente de determinação de regressão linear múltipla é definido de forma análoga ao já apresentado na regressão simples.

De notar que para $k \geq 2$ já não temos o quadrado de nenhum coeficiente de correlação como acontecia na regressão simples. No entanto, o coeficiente de determinação pode continuar a ser pensado como uma medida da quantidade de variabilidade explicada (ou tida em conta) pelo modelo de regressão já que consiste na razão entre a soma dos quadrados devido aos resíduos e a soma dos quadrados total. Tal como anteriormente, um bom ajuste deve refletir num valor de $R^2$ próximo de 1.

Atenção que este coeficiente pode induzir em erro. Ao adicionarem-se variáveis (regressoras) ao modelo, estamos sempre a aumentar o valor de $R^2$ e nem sempre essas variáveis são estatisticamente significativas. Assim, modelos com um valor elevado de $R^2$ podem produzir más estimativas da resposta média $E[Y_i]$.

Tal como na regressão linear simples, define-se o coeficiente de determinação ajustado para corrigir o viés do coeficiente de determinação:

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{S_{YY}/(n-1)}. \qquad (3.113)$$

- Variância dos erros

Tal como na regressão simples podemos comparar o valor do estimador do desvio padrão dos erros $\hat{\sigma} = \sqrt{MS_E}$ com a estimativa do desvio padrão da mostra de observações $Y_i, S_{C_Y}$. Se a estimativa do desvio padrão dos erros não for significativamente inferior à estimativa do desvio padrão de $Y$ então o modelo de regressão não é melhor do que a simples média da amostra $Y_1,\ldots,Y_n$ para predizer a variável dependente. Também como acontecia na regressão linear simples, note-se que este procedimento é equivalente ao estudo do coeficiente de determinação ajustado apresentado na seção anterior já que

$$\left(\frac{\hat{\sigma}}{S_{C_Y}}\right)^2 = 1 - R^2_{adj} \ .$$

### 3.3.9 Adequação do modelo de regressão

#### *3.3.9.1 Análise de resíduos*

É importante realçar que os métodos para validar o modelo de regressão linear múltipla são basicamente os mesmos apresentados para a regressão linear simples. Importar realçar alguns aspectos adicionais.

1. Para averiguar se os erros têm distribuição normal é usual traçar um $QQ-plot$ ou um $PP-plot$ para os resíduos, com base na distribuição normal. Também se pode proceder a testes de ajustamento dos resíduos a uma distribuição normal.
2. Para averiguar se os erros são aleatórios (independentes) e se a variância é constante constroem-se gráficos de resíduos versus valores preditos $\hat{Y}_i$ (ou versus valores observados, ou versus valores das regressoras) que devem apresentar uma mancha de pontos aleatórios com o mesmo tipo de dispersão em torno do eixo das abcissas.

O número de observações disponíveis para a análise de regressão deve ser no mínimo 3 a 4 vezes maior que o número de coeficientes da equação de regressão que serão estimados. Esta regra procura evitar um falso ajuste causado pelas oscilações que podem ocorrer nas variáveis independentes e que são de difícil detecção nas amostras muito pequenas (CPRM, 2015).

### 3.3.10 Seleção de variáveis na regressão múltipla

Muitas vezes quando dispomos de observações $Y_i$, $i = 1, \ldots, n$, acompanhadas de vários valores correspondentes a várias regressoras coloca-se a questão: Todas as regressoras são importantes para o modelo? Pode acontecer que uma ou mais regressoras não contribuam para um melhor ajuste do modelo e como tal não vale a pena considerá-las. Não esquecer que a inclusão de regressoras faz diminuir o número de graus de liberdade de várias estatísticas e como tal aumenta a incerteza de alguns dos resultados. Existem diversos procedimentos para selecionar um conjunto de variáveis (regressoras) tidas como fundamentais ou mais importantes em cada problema.

**Backward** - Neste método começa-se com todas as variáveis e eliminam-se as menos importantes uma a uma. Começa-se por calcular os valores das estatísticas *F* para cada uma das regressoras, ou seja $T_i^2, i = 1,\ldots,n$ com $T_0^2$ dado pela Equacão (3.30). Com base nas estatísticas de teste $F$ com valores inferiores a um certo limiar pré-estabelecido, retira-se a variável com menor valor de $F$. Voltam-se a calcular os valores das estatísticas $F$ com base no novo conjunto de regressoras, e repete-se o procedimento até que nenhuma estatística $F$ seja inferior ao valor pré-estabelecido.

**Forward** - Neste método as variáveis são introduzidas uma a uma. A primeira a ser introduzida é aquela que tiver maior coeficiente de correlação (em módulo) com a variável dependente $Y$. Nos passos seguintes são introduzidas de forma sequencial as variáveis com maior coeficiente de correlação parcial entre a variável dependente e a variável que se pretende introduzir tendo em conta as variáveis já introduzidas. O coeficiente de correlação parcial entre $Y$ e $x_j$, dado que $x_{i_1},\ldots,x_{i_l}$ já estão no modelo, é o coeficiente de correlação entre

1. os resíduos obtidos da regressão de $Y$ como função de $x_{i_1},\ldots,x_{i_l}$, e
2. os resíduos obtidos da regressão de $x_j$ com função de $x_{i_1},\ldots,x_{i_l}$.

Em cada passo é avaliado o valor da estatística de teste correspondente ao novo parâmetro de regressão introduzido (este tipo de valor é muitas vezes designado *partial F* ou *F to enter*). Se a estatística teste da distribuição $F$ for inferior a um determinado valor pré-estabelecido a variável que se acabou de introduzir é eliminada e considera-se uma nova variável (enquanto as houver).

**Stepwise** - Este método combina os anteriores. Basicamente é um procedimento *forward* pois vai adicionando variáveis uma a uma. No entanto, em cada passo é feita uma análise das variáveis já introduzidas até aí, por forma a garantir que permanecem relevantes após a introdução da nova variável. Este é o método mais completo dos três apresentados.

Em situações onde as variáveis explicativas são fortemente correlacionadas podem ocorrer problemas na regressão múltipla. Variáveis colineares não fornecem novas informações, dificultando a interpretação dos coeficientes obtidos na regressão, pois em alguns casos o sinal do coeficiente de regressão pode ser o oposto do esperado. Por isso

é fortemente recomendável a montagem de uma matriz de coeficientes de correlação simples entre as variáveis explicativas para verificar a existência de uma possível colinearidade entre essas variáveis ou uma possível multicolinearidade entre uma e as outras varíáveis. Um modo expedito de evitar a colinearidade é a eliminação de uma, entre cada conjunto de duas variáveis explicativas que apresentarem coeficientes de correlação superiores a 0,85 (ver CPRM, 2015, p. 389). Desse modo, espera-se que as variáveis mantidas no modelo de regressão contribuam significativamente para explicar a variabilidade de $Y$ .

# 4 Desenho Metodológico

## 4.1 Introdução

A metodologia descreve como a pesquisa foi realizada para que os objetivos propostos pudessem ser respondidos. Desta forma é apresentado o delineamento do estudo, a população estudada, o plano de recolha dos dados e como os dados foram analisados.

## 4.2 Delineamento do Estudo

Trata-se de um estudo de pesquisa documental uma vez que a fonte de recolha de dados está restrita a documentos e/ou relatórios disponíveis essencialmente nos *sites* oficiais de organizações internacionais de saúde e de governos. Associado à pesquisa bibliográfica, utilizou-se o desenho ecológico ou estatístico tendo em conta a natureza do estudo e a abrangência deste tipo de desenho. Eis a sua descrição:

Nos estudos ecológicos a unidade de análise são populações ou grupos de pessoas pertencentes a uma área geográfica definida e caracterizam-se por proporcionarem um proveitoso início para as pesquisas epidemiológicas mais detalhadas (Beaglehole et al., 2001 citado por Gesser, 2005). Nestes estudos procura-se avaliar como os contextos social e ambiental podem afetar a saúde de grupos populacionais.

O termo estudo ecológico tem origem no uso de áreas geográficas como unidades de análise e, por extensão, generalizou-se para outras situações em que a unidade é formada por um grupo (Pereira, 2001 *apud* Gesser, 2005).

A possibilidade de se trabalhar com dados de populações com características amplamente diferentes é um forte atrativo ao uso do estudo ecológico (Beaglehole et al., 2001 *apud* Gesser, 2005), justificando assim a escolha desse desenho de estudo para este trabalho.

Segundo Pereira (2001) citado por Gesser, (2005), uma variável ecológica é aquela que descreve o que ocorre em grupos de indivíduos, como por exemplo, o percentual de analfabetos. Os dados já estão agrupados e não se sabe se um determinado indivíduo tem esta ou aquela característica.

O estudo tem uma abordagem quantitativa, em virtude de todos os dados serem numéricos e por terem sido utilizadas técnicas estatísticas descritivas e analíticas.

## 4.3 Métodos de abordagem

Os tipos de métodos considerados são o método **indutivo**, onde a partir de uma análise de dados particulares, se encaminha para noções gerais e o método **hipotético-dedutivo**, onde se estabelece um problema, a colocação de hipótese ou soluções provisórias e a tentativa de resolução deste problema.

## 4.4 População e amostra estudada

A população de estudo foi constituída por todos os agregados familiares que residiam em habitações não coletivas, excluindo portanto os residentes em instituições coletivas (hotéis, prisões, hospitais, quartéis, etc.) existentes durante a realização do IBEP entre 2008 e 2009. A amostra do estudo foi determinada com base em 11.852 agregados a nível nacional, destes cerca de 658 correspondendo a cada uma das províncias do país, com excepção de Luanda, onde o tamanho da amostra é de 1.392.

## 4.5 A Base de Dados

O base de dados engloba todos as comunas angolanas, dispensando desta forma os métodos estatísticos de cálculo de tamanho da amostra ou de métodos de amostragem para recolha da mesma.

Assim sendo, o estudo tem sua unidade de pesquisa retirada do universo das comunas angolanas entre 2008 e 2009, facto que vem caracterizar o desenho do estudo como sendo do tipo ecológico.

O *software* Microsoft Excel foi escolhido como plataforma para acolher os dados, em virtude de sua versatilidade em armazenamento e conversão de dados, além de sua disponibilidade nos computadores pessoais.

As estatísticas descritivas, os cálculos dos coeficientes de correlação e a análise de regressão, que formulou os modelos explicativos da esperança de vida das populações, foram realizados recorrendo ao SPSS 20.0.0 *for Windows*, após a conversão do banco de dados formato ".xlsx" em um banco de dados ".sav".

A exportação dos dados para o *software* SPSS 20.0.0 *for Windows*, se fez necessária pelas limitações do Microsoft Excel em análises estatísticas mais detalhadas.

### 4.5.1 Composição da Base de Dados

A base de dados contém a identificação da província pelo nome e obviamente os dados das variáveis de estudo.

As variáveis em estudo são compostas por dados numéricos que expressam as características das províncias angolanas com relação à esperança de vida em anos, a taxa de mortalidade infantil até um ano de vida, o rendimento *per capita*, , o número de médicos residentes por mil habitantes, o número de pessoas com acesso à água potável, o tamanho da população, o número de camas hospitalares, a população, a mortalidade por malária, a mortalidade por VIH/SIDA e o acesso à educação.

As variáveis em estudo podem ser classificadas em dependentes e independentes (Larson & Farber, 2004). A esperança de vida é a dependente, pois o estudo vai procurar o seu grau de dependência com as demais variáveis em estudo, denominadas independentes.

O código da província, contido na base de dados, é o identificador utilizado pelo INE. Também serão inseridos o nome da província. Este procedimento é útil aquando da elaboração das tabelas estatísticas descritivas que refletirão as províncias com melhores e piores condições em cada atributo.

A esperança de vida ou variável-alvo das análises estatísticas (variável dependente do modelo de regressão) representa o número médio de anos que as pessoas viveriam a partir do seu nascimento.

A variável independente mortalidade infantil tem o propósito de expressar o número de crianças que não sobrevivem ao primeiro ano de vida em cada mil crianças nascidas vivas.

O rendimento *per capita* é entendido como a razão entre o somatório do rendimentorendimento *per capita* de todos os indivíduos e o número total desses indivíduos. O rendimento *per capita* de cada indivíduo é definido como a soma dos salários de toda população dividida pelo número de habitantes. Valores expressos em kwanzas entre Maio de 2008 e Maio 2009.

O percentual de pessoas com água potável é o percentual de pessoas que vivem em domicílios com água canalizada para um ou mais cômodos, proveniente de rede geral,

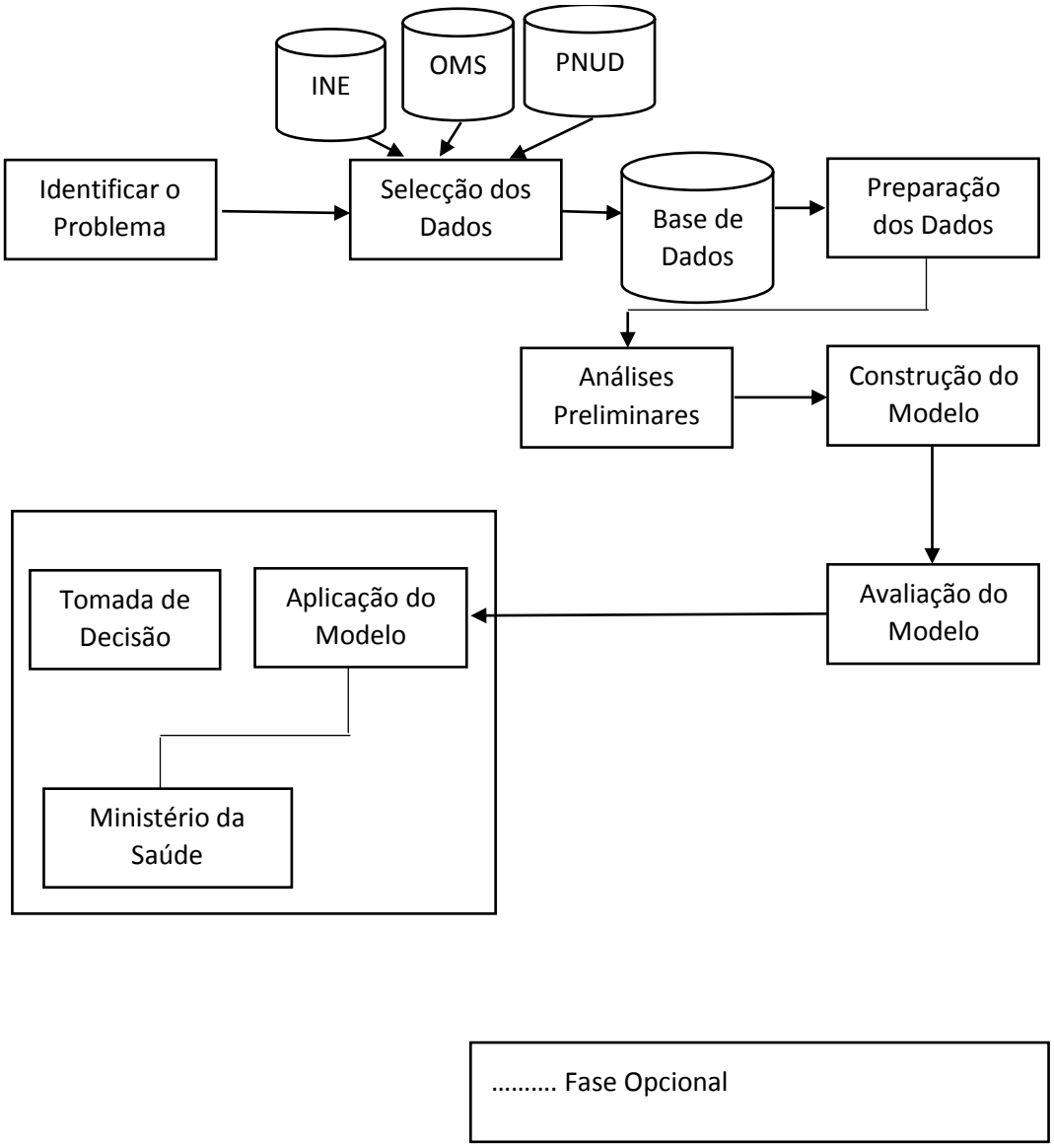
de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-cisterna.

A taxa de alfabetização é o percentual das pessoas acima de 15 anos de idade que são alfabetizados, ou seja, que sabem ler e escrever pelo menos um bilhete simples. Essa variável é um dos indicadores componentes do IDH-Educação (dimensão do Índice de Desenvolvimento Humano). A taxa de alfabetização participa com dois terços do peso do indicador IDH-Educação, caracterizando assim, como sendo o componente mais importante.

## 4.6 Métodos de Análise de Dados

Foi adotado como critério metodológico de análise de dados aquele proposto por Berry e Linoff (1997) citado por Gesser, 2005. Essa metodologia, apresentada na Figura 4.1 demonstra os passos a serem empregues no estudo, visando a geração de um modelo explicativo para a esperança de vida, além da extração de conhecimento sobre a base de dados.

**Figura 4.1** Modelo do estudo para análise da base de dados




**Fonte:** Adaptado de Gesser (2005).

Os dados das províncias angolanas com registos retirados dos sítios Web do INE (2011), da WHO (2010) das Nações Unidas, foram agrupados em uma plataforma de base de dados em Excel que reúne em uma única folha de cálculo todos os registros númericos das variáveis preditores por hipótese da pesquisa. Em seguida estes dados foram preparados para análises estatísticas, exportando para o efeito a formatos compativeis com o software SPSS 20.0.0 *for windows.*

Para a construção do modelo explicativo, foram feitas análises estatísticas, das quais se destacam a correlação e a análise de regressão linear múltipla.

E entre as propriedades do coeficiente de correlação $R$ , pode-se destacar o facto de que seu valor é um número adimensional. Como foi dito anteriormente é um estimador do correspondente parâmetro $\rho$. O seu sinal pode ser positivo ou negativo e sua faixa de variação está compreendida entre $-1$ e $1$. O coeficiente de correlação indica o grau da relação linear obtida, ou o grau de ajuste de uma recta ao conjunto dos pontos da amostra.

Quanto mais próximo $R$ estiver de $1$, mais próximos estarão os pontos de ajuste integral a uma recta crescente. Quanto mais próximo $R$ estiver de $-1$ , mais próximos estarão os pontos de ajuste integral a uma recta decrescente. Se $R=0$, não foi identificada relação numérica linear para os pares de valores de amostra analisada (Bruni, 2011).

Para se descobrir se a correlação encontrada na amostra de dados também ocorre na população, testam-se as hipóteses de existir ou não esta correlação na população, definidas em (3.65), através do teste de significância sobre o coeficiente de correlação de Pearson.

A análise de regressão, como referido no presente trabalho, fornece uma função matemática que descreve a relação entre duas ou mais variáveis. A natureza da relação é caracterizada por esta função ou equação de regressão. Esta equação pode ser usada para estimar ou predizer valores futuros de uma variável, com base em valores conhecidos ou supostos, de uma ou mais variáveis relacionadas. Logo, a análise de regressão pode servir como um instrumento explicativo para uma série de questões (Gesser, 2005 apud Anderson et al., 2003), como por exemplo: "Porque não somos todos da mesma altura?", "Porque não temos o mesmo rendimento familiar", "Porque algumas populações têm esperança de vida maior do que outras?"

Tal como já foi referido no capítulo anterior, a análise de regressão baseia-se na elaboração de uma função matemática do tipo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Todos os aspectos inerentes a esta equação já foram discutidos no capítulo 3.

# 5 Aplicação da Análise de Correlação e Regressão Linear Múltipla

## 5.1 Correlação da esperança de vida com as demais variáveis em estudo

A correlação e a significância da esperança de vida com as variáveis independentes, rendimento *per capita*, taxa de mortalidade por VIH/SIDA, taxa de mortalidade infantil, taxa de mortalidade por malária, número de camas hospitalares, percentual de pessoas que vivem em casas com água canalizada, número de médicos residentes por mil habitantes, taxa de alfabetização e tamanho da população estão apresentada na Tabela 5.1.

**Tabela 5.1** Coeficiente de Correlação de Pearson entre a esperança de vida e as variáveis em estudo em Angola

**Correlations**

| | | Esperança de vida | Malaria | Água | Médicos | HIV/SIDA | Redimento | Mortalidade Infantil | População | Alfabetização | Camas |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pearson Correlation | Esperança de vida | 1,000 | -,614 | ,072 | ,526 | ,449 | ,530 | -,147 | ,439 | ,570 | -,242 |
| Sig. (2-tailed) | Esperança de vida | . | ,003 | ,389 | ,013 | ,031 | ,012 | ,280 | ,034 | ,007 | ,166 |

A partir destes resultados podemos notar que a variável mortalidade por VIH/SIDA apresenta correlação linear fraca positiva[2] $(R = 0,449)$, a variável acesso à água potável fraca positiva $(R = 0,072)$, a mortalidade por malária moderada negativa $(R = -0,614)$, a variável número de médicos moderada positiva $(R = 0,526)$, a variável rendimento moderada positiva $(R = 0,530)$, a mortalidade infantil fraca negativa $(R = -0,147)$, a população moderada positiva $(R = 0,439)$, a alfabetização moderada positiva $(R = 0,570)$ e a variável número de camas hospitalares apresenta correlação linear fraca negativa $(R = -0,242)$.

Na última linha da Tabela 5.1 foram apresentadas os $p - values$ de cada uma das variáveis quanto a todas as províncias de Angola, notando-se que as variáveis

[2] O critério utilizado para classificar o grau de intensidade da correlação linear baseia-se em Maroco (2007).

mortalidade infantil, camas e acesso à água potável possuiem correlação não significante com a variável dependente ($sig. > 0,05$), esperança de vida ao nascer em Angola.

A correlação linear indica se uma determinada variável independente caminha ou não no mesmo sentido com a variável independente. É uma condição necessária mas não suficiente, ou seja não permite identificar relação de causa e efeito entre as variáveis preditoras e a variável resposta (esperança de vida ao nascer).

## 5.2 Análise de Regressão Linear Múltipla

A seguir são apresentadas a análise de regressão linear múltipla com a respectiva equação do modelo que visa explicar a esperança de vida ao nascer, em Angola, com as diversas variáveis independentes (socioeconómicas).

**Tabela 5.2** Estatística Descritiva

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Esperança de vida | 48,67 | 2,701 | 18 |
| Mortes por malaria | 619,6711 | 610,88719 | 18 |
| Acesso à água | 461,9089 | 455,51821 | 18 |
| Número de médicos | 84,4506 | 37,62694 | 18 |
| Mortes por VIH/SIDA | 20,9233 | 22,50110 | 18 |
| Rendimentos | 473,94 | 192,844 | 18 |
| Mortalidade infantil | 192,94 | 3,298 | 18 |
| Populacional | 999557,33 | 1124204,252 | 18 |
| Alfabetização | 998,94 | 642,908 | 18 |
| Camas hospitalares | 728,4983 | 372,04125 | 18 |

O valor médio da esperança de vida ao nascer em Angola em anos é de aproximadamente de 48,67, da mortalidade por malária é de 619,6711 mortes/1000000 de habitantes, do número de habitantes a viver em domicílios com água canalizada é de 461,9089/1000000 de habitantes, do número de médicos é de 84,4506 médicos/1000000 de habitantes, da mortalidade/1000000 de habitantes por VIH/SIDA é de 20,9233 óbitos/1000000 de habitantes, do rendimento é de 473,94 kwanzas em média por dia, da mortalidade infantil é de 192,94 óbitos/1000 nados vivos, da popolução é de 999557,33 habitantes, da alfabetização é 998,4983 alfabetizados /100000 habitantes e das camas hospitalares é de 728,4983 leitos/100000.

**Tabela 5.3** Sumário do Modelo

**Model Summary[c]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | ,614[a] | ,377 | ,338 | 2,198 | |
| 2 | ,777[b] | ,604 | ,551 | 1,810 | 2,419 |

a. Predictors: (Constant), Mortalidade por Malaria

b. Predictors: (Constant), Mortalidade por Malaria, Rendimento

c. Dependent Variable: Esperança de vida

Observando a Tabela 5.3, concluímos que $R^2$ e $R^2_{adj}$ tomam valores aproximados, sendo que o maior valor de $R^2_{adj}$ corresponde ao modelo em que são consideradas as duas variáveis explicativas, mortalidade por malária e o rendimento. Concluímos que este modelo será provavelmente o que melhor explica os valores da esperança de vida ao nascer em Angola. Este valor permite-nos afirmar que 55,1% $\left(R^2_{adj} = 0,551\right)$ da variabilidade da esperança de vida em Angola é explicada por este modelo.

**Tabela 5.4 Tabela da ANOVA**

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 46,732 | 1 | 46,732 | 9,677 | ,007[b] |
| | Residual | 77,268 | 16 | 4,829 | | |
| | Total | 124,000 | 17 | | | |
| 2 | Regression | 74,844 | 2 | 37,422 | 11,419 | ,001[c] |
| | Residual | 49,156 | 15 | 3,277 | | |
| | Total | 124,000 | 17 | | | |

a. Dependent Variable: Esperança de vida

b. Predictors: (Constant), Mortalidade por Malaria

c. Predictors: (Constant), Mortalidade por Malaria, Rendimento

Mediante a análise do valor de significância do teste $F$ $sig. = 0,001$ concluímos que o modelo é altamente significativo. Verifica-se que a esperança de vida é explicada pelas duas variáveis independentes (Mortalidade por Malaria e Rendimento). Este fato é

confirmado pelo valor de significância do teste $t$ da Tabela 5.5 ($sig. = 0,03$, $sig. = 0,01$, respectivamente).

**Tabela 5.5** Coeficientes

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95,0% Confidence Interval for B | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound | Tolerance | VIF |
| 1 | (Constant) | 50,349 | ,749 | | 67,246 | ,000 | 48,761 | 51,936 | | |
| | MMalaria | -,003 | ,001 | -,614 | -3,111 | ,007 | -,005 | -,001 | 1,000 | 1,000 |
| 2 | (Constant) | 47,056 | 1,282 | | 36,697 | ,000 | 44,323 | 49,789 | | |
| | MMalaria | -,003 | ,001 | -,570 | -3,494 | ,003 | -,004 | -,001 | ,992 | 1,008 |
| | Receitas | ,007 | ,002 | ,478 | 2,929 | ,010 | ,002 | ,012 | ,992 | 1,008 |

a. Dependent Variable: LifeExp

Também da análise da Tabela 5.5 concluimos que o modelo explicativo tem como equação

$$\boldsymbol{Y = 47,056 - 0,003 \times \text{Mortalidade por malária} + 0,007 \times \text{Rendimento médio per capita}} \quad (5.1)$$

A variável mortalidade por malária apresenta coeficiente negativo o que faz sentido, uma vez que, o aumento da mortalidade por malária faz diminuir a esperança de vida ao nascer. No entanto, a variável rendimento *per capita* apresenta coeficiente positivo, o que parece fazer sentido, uma vez que, o aumento da receita média faz com que os níveis de pobreza diminuam e pode fazer com a esperança de vida aumente.

**Tabela 5.6** Variáveis Excluídas

**Excluded Variables[a]**

| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics |
|---|---|---|---|---|---|---|
| | | | | | | Tolerance |
| 1 | Acesso à água | ,142[b] | ,702 | ,493 | ,178 | ,988 |
| | Número de médico | ,354[b] | 1,792 | ,093 | ,420 | ,877 |
| | Mortes por VIH/SIDA | ,292[b] | 1,459 | ,165 | ,353 | ,911 |
| | Rendimento | ,478[b] | 2,929 | ,010 | ,603 | ,992 |
| | Mortalidade infantil | -,011[b] | -,053 | ,958 | -,014 | ,950 |
| | População | ,359[b] | 1,954 | ,070 | ,451 | ,980 |
| | Taxa de alfabetização | ,466[b] | 2,748 | ,015 | ,579 | ,960 |
| | Camas hospitalares | -,157[b] | -,775 | ,450 | -,196 | ,979 |
| 2 | Acesso à água | ,092[c] | ,545 | ,595 | ,144 | ,976 |
| | Número de médico | ,189[c] | 1,012 | ,329 | ,261 | ,753 |
| | Mortes por VIH/SIDA | ,088[c] | ,452 | ,658 | ,120 | ,732 |
| | Mortalidade infantil | ,094[c] | ,538 | ,599 | ,142 | ,910 |
| | População | -,132[c] | -,435 | ,670 | -,115 | ,302 |
| | Taxa de alfabetização | ,051[c] | ,093 | ,927 | ,025 | ,095 |
| | Camas hospitalares | -,073[c] | -,427 | ,676 | -,113 | ,948 |

a. Dependent Variable: Esperança de vida

b. Predictors in the Model: (Constant), Mortalidade por Malaria

c. Predictors in the Model: (Constant), Mortalidae por Malaria, Receitas

Das variáveis excluídas do segundo modelo tenho a salientar o facto das variáveis independentes acesso à água potável e número de médicos possuírem coeficientes positivos apesar de não serem significativos em relação ao modelo gerado.

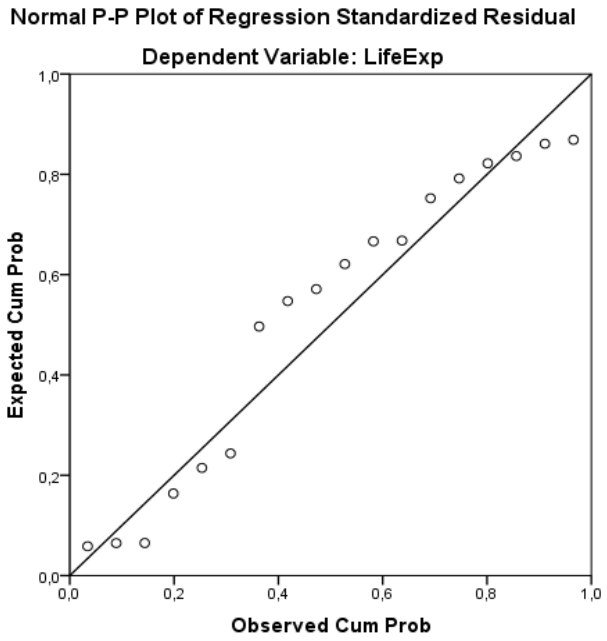
### 5.2.1 Verificação dos pressupostos do modelo

Com o objetivo de verificar se o modelo (5.1) é adequado, de seguida é feita uma análise dos pressupostos da regressão linear múltipla.

- NORMALIDADE DOS RESÍDUOS

Observando a Gráfico 5.1 parecem existir alguns pontos que se afastam da diagonal principal, não sendo conclusivos quanto à normalidade dos resíduos.

**Gráfico 5.1** Normal p-p plot da regressão dos resíduos estandardizados




Para confirmar a normalidade realizámos o teste K-S apresentado na Tabela 5.7. Mediante o valor de significância obtido (0,667) confirmamos que os resíduos são normalmente distribuídos, devido a não se rejeitar a hipótese nula.

**Tabela 5.7** Teste K-S

**One-Sample Kolmogorov-Smirnov Test**

| | | Unstandardized Residual |
|---|---|---|
| N | | 18 |
| Normal Parametersa,b | Mean | 0E-7 |
| | Std. Deviation | 1,70044403 |
| Most Extreme Differences | Absolute | ,163 |
| | Positive | ,116 |
| | Negative | -,163 |
| Kolmogorov-Smirnov Z | | ,692 |
| Asymp. Sig. (2-tailed) | | ,725 |
| Exact Sig. (2-tailed) | | ,667 |
| Point Probability | | ,000 |

a. Test distribution is Normal.

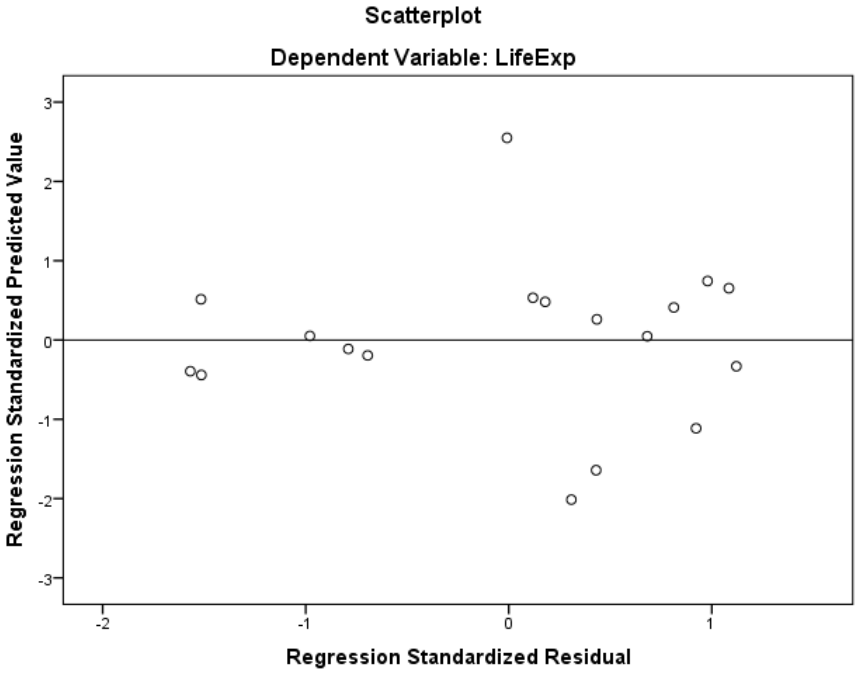
b. Calculated from data.

- AUTOCORRELAÇÃO DOS RESÍDUOS

Considerando o resultado obtido para o teste de Durbin-Watson, apresentado na Tabela 5.3 (2,419) e uma vez que esse valor pertence ao intervalo $[d_U, 4-d_U[$, concluímos que os resíduos são independentes (ver Seção 3.2.1.2.1 e Tabela do Anexo 2).

- HOMOSCEDASTICIDADE DOS RESÍDUOS

**Gráfico 5.2** Gráfico dos resíduos estandardizados




A partir da análise gráfica dos resíduos estandardizados, Gráfico 5.2 5.7, como os resíduos se distribuem aleatoriamente em torno de zero, concluímos que os resíduos são homoscedásticos. (ver Marôco (2011) e Hall, Neves, & Pereira (2011)).

- COLINEARIADADE

Uma vez que se trata de um modelo de regressão linear múltipla, um dos pressupostos a ser verificado é se existe colinearidade entre as duas variáveis independentes. Esta inferência é confirmada pelos pelos *Condition Index* na Tabela 5.8, já que estes valores são inferiores a 15 (ver Marôco, 2007).

**Tabela 5.8** Diagnóstico da Colinearidade

**Collinearity Diagnostics[a]**

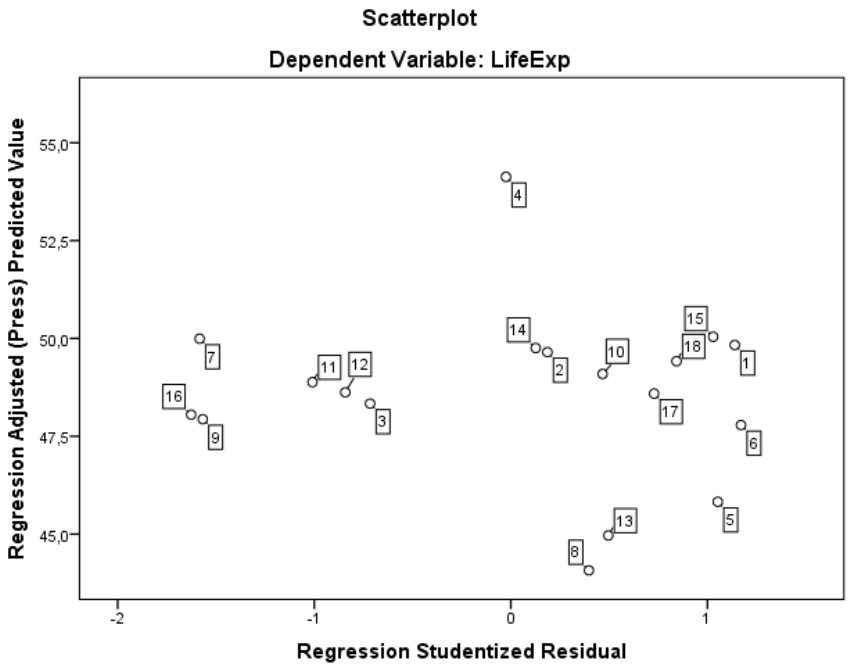
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | |
|---|---|---|---|---|---|---|
| | | | | (Constant) | Mortalidade por Malaria | Rendimentos |
| 1 | 1 | 1,722 | 1,000 | ,14 | ,14 | |
| | 2 | ,278 | 2,489 | ,86 | ,86 | |
| 2 | 1 | 2,540 | 1,000 | ,02 | ,05 | ,02 |
| | 2 | ,396 | 2,532 | ,02 | ,84 | ,07 |
| | 3 | ,064 | 6,310 | ,96 | ,11 | ,91 |

a. Dependent Variable: Esperança de vida

Tendo em conta que para as duas variáveis independentes os valores de $\text{VIF} < 5$, como podemos confirmar na Tabela 5.5, concluímos que não existe colinearidade entre as duas variáveis explicativas.

- OUTLIERS E OBSERVAÇÕES INFLUENTES

**Gráfico 5.3** Gráfico resíduos press




Pela análise gráfica dos resíduos press, Gráfico 5.3, temos que, não existem *Outliers*, dado que apresenta resíduos com valores absolutos não superiores a 1,96.

**Tabela 5.9** Estatística dos Resíduos

**Residuals Statistics[a]**

| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | 44,44 | 54,02 | 48,67 | 2,098 | 18 |
| Std. Predicted Value | -2,013 | 2,549 | ,000 | 1,000 | 18 |
| Standard Error of Predicted Value | ,435 | 1,696 | ,672 | ,316 | 18 |
| Adjusted Predicted Value | 44,07 | 54,13 | 48,61 | 2,202 | 18 |
| Residual | -2,839 | 2,029 | ,000 | 1,700 | 18 |
| Std. Residual | -1,568 | 1,121 | ,000 | ,939 | 18 |
| Stud. Residual | -1,626 | 1,171 | ,015 | ,988 | 18 |
| Deleted Residual | -3,053 | 2,214 | ,056 | 1,885 | 18 |
| Stud. Deleted Residual | -1,731 | 1,187 | -,002 | 1,015 | 18 |
| Mahal. Distance | ,038 | 13,981 | 1,889 | 3,358 | 18 |
| Cook's Distance | ,001 | ,111 | ,034 | ,029 | 18 |
| Centered Leverage Value | ,002 | ,822 | ,111 | ,198 | 18 |

a. Dependent Variable: Esperança de vida

A existência ou não de *Outliers* pode ser feita através do valor máximo de *Student Deleted Residual* (1,187<1,96). Facto não confirmado pelo Valor Centrado de Leverage $LEV = 0,822 > \frac{3(p+1)}{18} = \frac{9}{18} = 0,5$. Este valor mostra que poderão existir *Outliers* . Será que os possíveis valores são excessivamente influente? A fim de responder à esta questão temos que analisar o valor da Distância de Cook, isto é, $COOK = 0,111 < \frac{4}{n-p-1} = \frac{4}{15} = 0,67$. Portanto, a resposta é negativa. Em geral, considera-se que observações com Distância de Cook superior a 1 são excessivamente influentes (ver Maroco, 2007).

**Conclusão**

Todos pressupostos da análise de regressão linear foram validados. Temos o modelo (5.1) como o modelo válido, adequado aos dados.

# 6 Conclusões e Sugestões

Esta investigação teve como objectivo a elaboração de um modelo explicativo, gerado pela análise de regressão linear múltipla, para a esperança de vida, baseados em fatores socioeconómicos de Angola.

Para o suporte teórico do presente estudo foram levantados na literatura informações referentes às variáveis socioeconómicas que se apresentam mais frequentemente relacionadas com a esperança de vida populacional, permitindo o alcance dos objetivos da pesquisa.

## 6.1 Conclusões

Os objetivos propostos da pesquisa foram respondidos através da análise estatística das variáveis que compõem o Inquérito Sobre o Bem-Estar da População IBEP de 2010.Os dois primeiros objectivos do estudo foram alcançados.

O último objetivo específico foi respondido com a aplicação da correlação e da análise de regressão à base de dados, sendo identificadas quais as variáveis independentes em estudo justificam a variável dependente, esperança de vida. A correlação linear de Pearson foi o método escolhido para se observar a correlação existente entre a variável dependente e as independentes. As variáveis mortalidade por VIH/SIDA, número de médicos residentes, rendimento *per capita* e taxa de alfabetização foram as que se apresentaram mais moderadamente correlacionada com a esperança de vida em todo pais.

Desta forma, entende-se que as políticas públicas priorizem as questões de redução dos índices de mortalidade por malária e mortalidade infantil. Desta forma estarão também agindo na elevação da esperança de vida populacional.

O modelo gerado apresentou coeficiente de determinação ajustado ( $R^2_{adj} \geq 0,551$). Tal fato, segundo a literatura, significa que a qualidade do ajuste não é tão forte. Contudo os coeficientes confirmam a relevância que os determinantes sociais têm na esperança de vida populacional.

O presente relatório mostrou que a esperança de vida aos nascer em Angola sofre uma forte influência das variáveis socioeconômicas e com particular realce as de saúde pública. Para a promoção e criação de ambientes saudáveis torna-se necessário compreender os fenómenos relacionados às variáveis socioeconómicas.

A reorientação dos serviços de saúde na direção da concepção da promoção da saúde, além do provimento de serviços assistenciais, está entre as medidas preconizadas na Carta de Ottawa. Ainda segundo a mesma carta a promoção da saúde vai para além dos cuidados de saúde. Ela coloca a saúde na agenda de prioridades dos políticos e dirigentes em todos os níveis e setores, chamando-lhes a atenção para as consequências que suas decisões podem ter no campo da saúde.

Os modelos de regressão linear obtidos para cada uma das regiões angolanas mostraram claramente a superação do modelo biomédico, centrado na doença como fenômeno individual e na assistência médica curativa desenvolvida nos estabelecimentos médico-assistenciais como foco essencial da intervenção. Logo, são necessárias profundas transformações na organização e financiamento dos sistemas e serviços de saúde, assim como nas práticas, nas infraestruturas e na formação dos profissionais. Ademais, os serviços de saúde baseados na cura das enfermidades é arcaico. Tem que se passar pela prevenção e melhoria da qualidade de vida das populações.

Notamos que as observações para cada uma das variáveis foram 2 vezes ao número das mesmas, o que certa forma limita o estudo. Acreditamos que se considerássemos as regiões por municípios ou comunas, os resultados seriam mais abrangentes.

Um outro dado a reter é o nível acentuado de pobreza da população angolana. Portanto, torna-se necessário que as estruturas centrais levem acabo políticas sociais que visem mitigar esta problemática.

Com o modelo gerado torna-se possível ao gestor público conhecer de forma antecipada o impacto que uma eventual alteração em alguma das variáveis independentes causa no valor da esperança de vida da província. A utilidade do modelo se dará em futuras aplicações como instrumento auxiliar a elaboração de políticas públicas de saúde voltadas para o incremento da esperança de vida das pessoas.

## 6.2 Sugestões

É importante realçar que os resultados alcançados não estão finalizados, uma vez que nem todas as variáveis ou fatores determinantes foram utilizados nesta pesquisa.

Por outro lado, os métodos empregues na pesquisa foram quantitativos, o que pressupõe que só os fatores relevantes no estudo é que são descritos, fechando a porta a outros métodos que envolvam contacto direto com as populações e outros intervenientes, em

suma a sociedade. Assim, em estudos futuros dever-se-á optar por abordagens qualitativas onde todas as forças da sociedade tenham a possibilidade de participar, dando a sua opinião.

Após a análise elaborada neste trabalho como sugestões temos a salientar o seguinte:

1. Que haja uma maior desconcentração na gestão dos recursos destinados a prestação dos serviços de saúde pública e combate da pobreza;
2. Que se criem condições sociais que permitam a colocação de especialistas em saúde pública em regiões recônditas;
3. Que haja uma maior iteração entre as academias e os serviços de saúde visando a implementação de uma pesquisa mais profunda e eficaz no sentido de identificar outros fatores de âmbito cultural que estão por detrás da baixa esperança de vida ao nascer em Angola;
4. Que haja publicação por parte dos órgãos competentes de todos os indicadores socioeconómicos no sentido de facilitar o trabalho dos investigadores.

## 6.3 Limitações do estudo

A pesquisa em causa apresenta limitações, uma vez que o número de observações das variáveis contidas na base de dados não são os mais desejados já que referem-se apenas as dezoitos províncias de Angola.

## 6.4 Sugestões para trabalhos futuros

Assim sendo, em trabalhos futuros sugerimos que se considere um número de observações maior, por exemplo considerando os municípios de cada província; o que aumentaria o número de observações e consequentemente um estudo mais aprofundado dos reais fatores que influenciam a esperança de vida em Angola.

# 7 Bibliografia


Action, P. (13 de Fevereiro de 2015). *Portal Action*. Obtido de Portal Action: http://www.portalaction.com.br/content/27-an%C3%A1lise-de-res%C3%ADduos

Agency, C. I. (s.d.). *The World Factbook*. Obtido em 22 de Outubro de 2013, de The World Factbook: https://www.cia.gov/library/publications/the-world-factbook/

Angola, R. d. (5 de Fevereiro de 2010). Diário da República. *Diário da República*. Luanda, Luanda, Angola: Imprensa Nacional - E.P.

Bruni, A. L. (2011). *Estatística Aplicada à Gestão Empresarial.* São Paulo: Editora Atlas.

Bruni, A. L. (2011). *PASW Aplicado à Pesquisa Académica.* SÃO PAULO: Editora Atlas.

Calado, V., & Montegomery, D. C. (2003). *Planejamento de Experimento Usando o Statistica.* Rio de Janeiro: E-Papers.

Cosep, Consaúde, & International, M. (2011). *Inquérito de Indicadores de Malária em Angola 2011*. Calverton - Maryland: Macro International.

CPRM. (3 de Março de 2015). *CPRM-Companhia de Pesquisa de Recursos Mineiros.* Obtido de CPRM-Serviços Geológicos do Brasil: http://www.cprm.gov.br/publique/media/cap9-correl_regres.pdf

Gentle, J. E. (2009). *Computational Statistics.* New York: Springer.

Gesser, H. C. (2005). Obtido em 25 de Outubro de 2013, de https://repositorio.ufsc.br/bitstream/handle/123456789/101707/221720.pdf?sequence=1

Hall, A., Neves, C., & Pereira, A. (2011). *Grande Maratona de Estatística no SPSS.* Lisboa: Escolar Editora.

Hinkelmann, K., & Kempthorne, O. (2008). *Design and Analysis of Experiments.* New Jersey: John Wiley & Sons, Inc. .

Hinkelmann, K., & Kempthorne, O. (2008). *Design and Analysis os Experiments* (2ª Edição ed., Vol. 1). Virginia: John Wiley & Sons, Inc.

INE. (2010). *Inquérito Integrado sobre o Bem-Estar da População | IBEP.* Luanda - Angola: INE.

INE. (2012). *Projeção da População 2009-2015.* Luanda-Angola: INE.

INE. (2013). *Inquérito de Indicadores Básicos de Bem-Estar | QUIBB* . Luanda - Angola: Instituto Nacional de Estatística.

Larson, R., & Farber, B. (2004). *Estatística Aplicada*. São Paulo: Pearson Education do Brasil.

Marconi, M. d., & Lakatos, E. M. (2003). *Fundamentos de Metodologia Científica.* São Paulo: EDITORA ATLAS S.A.

Maroco, J. (2007). *Análise Estatística com o SPSS Statistics.* Pero Pinheiro: Edições Sílabo.

Martinez, L. F., & Ferreira, A. I. (2008). *Análise de Dados com SPSS Primeiros Passos.* Lisboa: Escolar Editora.

MINSA. (2003). *Inquérito de Indicadores Múltiplos.* Luanda-Angola: UNICEF.

MINSA. (2014). *Relatório de Progresso da Resposta Global à SIDA (GAPPR, 2014).* Luanda.

Montegomery, D. ( 2001). *Design and Analysis of Experiments* . Arizona - USA: Jonh Wilwy & Sons, Inc.

Montegomery, D. C., & Runger, G. C. (2011). *Applied Statistics and Probability for Engineers* (5 Edição ed.). Arizona : John Wiley & Sons, Inc.

MONTGOMERY, D. C. (2009). *Introduction to Statistical Quality Control.* USA: by John Wiley & Sons.

Morais, M. C. (2003). Estatística Computacional. Lisboa.

Nocedal, J., & Wright, S. J. (1999). *Numerical Optimization.* New York: Springer.

Nunes, C. F. (2012). *Probabilidades & Estatística.* Lisboa: Escolar Editora.

PNUD. (2006). *PNUD Brasil - Programa das Nações Unidas para o Desenvolvimento.* Obtido de PNUD Brasil - Programa das Nações Unidas para o Desenvolvimento: http://www.pnud.org.br/publicacoes/atlas_bh/release_longevidade.pdf

Rodrigues, S. C. (15 de Março de 2012). *Ubi Thesis - Conhecimento Online* . Obtido de Ubi Thesis Conhecimento Online : https://ubithesis.ubi.pt/bitstream/10400.6/1869/1/Tese%20Sandra%20Rodrigues.pdf

Saúde, 1. C. (25 de Fevereiro de 2015). *Direção Geral da Educação.* Obtido de Direção Geral da Educação: www.dgidc.min-edu.pt/educacaosaude/.../2_ottawa_nesase_semlogo.pdf

UNICEF. (2010). *Uma Angola melhor para TODAS as crianças.* Angola: UNICEF.

WHO. (2010). *Factsheets of Health Statistics.* Regional Office for Africa: WHO.

WHO. (2010). *WHO - World Health Organization*. Obtido em 22 de Outubro de 2013, de WHO - World Health Organization: http://www.who.int/en/

WHO. (2014). *World Health Statistics.* Genebra - Suiça: WHO.

# 8 Anexos

# Anexo 1

## Quadro 3.2.1 - Receitas médias mensais por pessoa, segundo os quintis (Kwanzas)

| | 1º quintil | 2º quintil | 3º quintil | 4º quintil | 5º quintil | Total* | Número de agregados |
|---|---|---|---|---|---|---|---|
| **Angola** | 1.414 | 3.023 | 5.086 | 8.288 | 26.035 | 8.767 | 8.530 |
| **Área de residência** | | | | | | | |
| Urbana | 1.689 | 3.912 | 6.485 | 10.550 | 32.784 | 11.077 | 4.563 |
| Rural | 1.230 | 2.436 | 3.804 | 6.032 | 16.383 | 5.967 | 3.967 |
| **Região (Total)** | | | | | | | |
| Luanda (capital do país) | 2.448 | 5.125 | 7.728 | 12.303 | 34.383 | 12.369 | 1.195 |
| Região Centro Sul | 1.051 | 2.231 | 3.727 | 6.663 | 23.571 | 7.435 | 1.563 |
| Região Este | 1.044 | 2.077 | 3.253 | 5.133 | 12.682 | 4.830 | 1.401 |
| Região Centro Norte | 1.500 | 2.849 | 4.183 | 6.627 | 19.795 | 6.972 | 1.480 |
| Região Sul | 1.809 | 3.708 | 5.709 | 8.421 | 26.331 | 9.187 | 1.407 |
| Região Norte | 1.471 | 2.695 | 4.166 | 6.419 | 18.822 | 6.711 | 1.484 |
| **Região e área de residência** | | | | | | | |
| Região Centro Sul | | | | | | | |
| Urbana | 1.268 | 3.228 | 6.321 | 11.297 | 39.740 | 12.322 | 803 |
| Rural | 1.014 | 1.988 | 3.023 | 4.965 | 12.488 | 4.676 | 760 |
| Região Este | | | | | | | |
| Urbana | 911 | 2.017 | 3.367 | 5.397 | 16.201 | 5.555 | 622 |
| Rural | 1.122 | 2.109 | 3.208 | 5.021 | 10.903 | 4.467 | 779 |
| Região Centro Norte | | | | | | | |
| Urbana | 1.195 | 2.705 | 4.249 | 7.424 | 23.709 | 7.844 | 726 |
| Rural | 1.697 | 2.896 | 4.142 | 6.248 | 17.321 | 6.429 | 754 |
| Região Sul | | | | | | | |
| Urbana | 1.749 | 3.378 | 5.327 | 8.145 | 26.309 | 8.847 | 753 |
| Rural | 1.961 | 4.046 | 6.006 | 8.697 | 26.766 | 9.477 | 654 |
| Região Norte | | | | | | | |
| Urbana | 1.430 | 3.017 | 4.912 | 8.151 | 24.574 | 8.329 | 618 |
| Rural | 1.507 | 2.611 | 3.898 | 5.704 | 15.533 | 5.840 | 866 |

**Capital do País** é formada pela província de Luanda.
**Região Centro Sul** é formada pelas províncias do Huambo, Bié, Benguela e Kwanza Sul.
**Região Este** é formada pelas províncias da Lunda Norte, Lunda Sul e Moxico e Kuando Kubango
**Região Centro Norte** é formada pelas províncias do Bengo, Malanje e Kwanza Norte.
**Região Sul** é formada pelas províncias do Namibe, Cunene e Huíla
**Região Norte** é formada pelas províncias de Cabinda, Uíge e Zaire.

***Indicador Chave do IBEP**: **Receitas médias mensais por pessoa**: é o quociente entre o valor das despesas totais do agregado e o respectivo número de membros no agregado.

# Anexo 2

**Tabela 3.3.2.1** Valores críticos do teste de Durbin-Watson

| | Nível de significância | Número de variáveis explicativas | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | | 4 | | 5 | |
| n | | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ |
| | 0,01 | 0,81 | 1,07 | 0,7 | 1,25 | 0,59 | 1,46 | 0,49 | 1,7 | 0,39 | 1,96 |
| 15 | 0,025 | 0,95 | 1,23 | 0,83 | 1,4 | 0,71 | 1,61 | 0,59 | 1,84 | 0,48 | 2,09 |
| | 0,05 | 1,08 | 1,36 | 0,95 | 1,54 | 0,82 | 1,75 | 0,69 | 1,97 | 0,56 | 2,21 |
| | 0,01 | 0,95 | 1,15 | 0,86 | 1,27 | 0,77 | 1,41 | 0,63 | 1,57 | 0,6 | 1,74 |
| 20 | 0,025 | 1,08 | 1,28 | 0,99 | 1,41 | 0,89 | 1,55 | 0,79 | 1,7 | 0,7 | 1,87 |
| | 0,05 | 1,2 | 1,41 | 1,1 | 1,54 | 1 | 1,68 | 0,9 | 1,83 | 0,79 | 1,99 |
| | 0,01 | 1,05 | 1,21 | 0,98 | 1,3 | 0,9 | 1,41 | 0,83 | 1,52 | 0,75 | 1,65 |
| 25 | 0,025 | 1,13 | 1,34 | 1,1 | 1,43 | 1,02 | 1,54 | 0,94 | 1,65 | 0,86 | 1,77 |
| | 0,05 | 1,2 | 1,45 | 1,21 | 1,55 | 1,12 | 1,66 | 1,04 | 1,77 | 0,95 | 1,89 |
| | 0,01 | 1,13 | 1,26 | 1,07 | 1,34 | 1,01 | 1,42 | 0,94 | 1,51 | 0,88 | 1,61 |
| 30 | 0,025 | 1,25 | 1,38 | 1,18 | 1,46 | 1,12 | 1,54 | 1,05 | 1,63 | 0,98 | 1,73 |
| | 0,05 | 1,35 | 1,49 | 1,28 | 1,57 | 1,21 | 1,65 | 1,14 | 1,74 | 1,07 | 1,83 |
| | 0,01 | 1,25 | 1,34 | 1,2 | 1,4 | 1,15 | 1,46 | 1,1 | 1,52 | 1,05 | 1,58 |
| 40 | 0,025 | 1,35 | 1,45 | 1,3 | 1,51 | 1,25 | 1,57 | 1,2 | 1,63 | 1,15 | 1,69 |
| | 0,05 | 1,44 | 1,54 | 1,39 | 1,6 | 1,34 | 1,66 | 1,29 | 1,72 | 1,23 | 1,79 |
| | 0,01 | 1,32 | 1,4 | 1,28 | 1,45 | 1,24 | 1,49 | 1,2 | 1,54 | 1,16 | 1,59 |
| 50 | 0,025 | 1,42 | 1,5 | 1,38 | 1,54 | 1,34 | 1,59 | 1,3 | 1,64 | 1,26 | 1,69 |
| | 0,05 | 1,5 | 1,59 | 1,46 | 1,63 | 1,42 | 1,67 | 1,38 | 1,72 | 1,34 | 1,7 |

**Fonte:** Portal Action (2015).

# Anexo 3

## Observações das variáveis de estudo

| Província | LifeExp | MMalária | AÁgua | NMédico | MHIV | Rendimento | InfMortality | População | TAlfabetização | NCamas |
|---|---|---|---|---|---|---|---|---|---|---|
| Cabinda | 52 | 126,7 | 539,76 | 103,9 | 73,49 | 493 | 194 | 394620 | 1239 | 894,53 |
| Zaire | 50 | 355,3 | 1133,59 | 95,88 | 2,82 | 525 | 199 | 354627 | 1058 | 741,62 |
| Uige | 47 | 760,69 | 221,12 | 37,03 | 3,17 | 466 | 190 | 945196 | 1077 | 978,63 |
| Luanda | 54 | 413,17 | 277,23 | 147,83 | 62,62 | 1195 | 190 | 5046323 | 3457 | 532,47 |
| Kuanza-Norte | 48 | 1652,67 | 1903,44 | 99,7 | 21,15 | 514 | 192 | 330979 | 734 | 601,25 |
| Kuanza-Sul | 50 | 553,07 | 132,64 | 79,25 | 4,17 | 345 | 194 | 1198758 | 763 | 776,64 |
| Malange | 47 | 200,42 | 595,15 | 70,38 | 45,9 | 477 | 192 | 653618 | 812 | 449,8 |
| Lunda-Norte | 45 | 2087,91 | 62,83 | 62,83 | 1,46 | 396 | 195 | 684417 | 666 | 1252,16 |
| Benguela | 45 | 830,22 | 111,24 | 47,51 | 2,32 | 415 | 194 | 1726057 | 857 | 1070,65 |
| Huambo | 50 | 70,67 | 160,04 | 52,65 | 4,85 | 349 | 185 | 1443388 | 915 | 1357,22 |
| Bié | 47 | 524,4 | 175,47 | 51,84 | 29,91 | 454 | 190 | 1003042 | 523 | 556,31 |
| Moxico | 47 | 285,99 | 415,81 | 105,47 | 14,2 | 313 | 195 | 493019 | 717 | 448,26 |
| Kuando Kubango | 46 | 1693,91 | 135,74 | 36,76 | 22,62 | 364 | 196 | 353619 | 610 | 520,33 |
| Namibe | 50 | 27,72 | 683,76 | 157,08 | 30,8 | 418 | 190 | 324673 | 973 | 1413,73 |
| Huila | 52 | 139,13 | 197,43 | 40,15 | 3,3 | 526 | 192 | 1818382 | 1033 | 197,98 |
| Cunene | 45 | 921,32 | 667,35 | 85,83 | 3,5 | 464 | 193 | 570918 | 1082 | 795,21 |
| Lunda-Sul | 50 | 192,95 | 295,27 | 122,78 | 43,85 | 328 | 198 | 342063 | 660 | 289,42 |
| Bengo | 51 | 317,84 | 606,49 | 123,24 | 6,49 | 489 | 194 | 308333 | 805 | 236,76 |

**Fonte:** INE (2011).