UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

FATORES EXPLICATIVOS NA OFERTA DE SERVIÇOS FARMACÊUTICOS EM PORTUGAL

MESTRADO EM MATEMÁTICA APLICADA À ECONOMIA E GESTÃO

Nádia Bachir

Trabalho de projeto orientado por:

Prof. Doutora Teresa Alpuim

Dr. José Pedro Guerreiro

2015

Um modelo é para ser usado, não é para acreditarmos nele.

O modelo não pretende ser mais do que uma caricatura da realidade.

Tem é de ser uma caricatura suficientemente boa,

para a reconhecermos sob aqueles traços simplificados.

Pestana e Velosa

Agradecimentos

Em primeiro lugar, agradeço a Deus por tantas coisas boas na minha vida e por ter escolhido para mim os melhores pais do Mundo que sempre me apoiaram em tudo, incentivaram-me a prosseguir os meus estudos e ensinaram-me que nada se consegue sem trabalho e dedicação.
À minha orientadora, Professora Teresa Alpuim, a minha especial gratidão pela atenção, confiança, e sobretudo pela disponibilidade e por me ter recebido sempre com um sorriso, mesmo quando tinha muito trabalho e eu aparecia com alguma dúvida. A sua ajuda foi indispensável na elaboração deste trabalho.

Este projeto surgiu de um contacto feito com a Associação Nacional de Farmácias por parte da professora Regina Bispo, que contribuiu na minha formação na área da Estatística, a quem agradeço muito não só o apoio como a amizade.
À ANF e ao Dr. José Pedro Guerreiro agradeço pelo interesse no projeto e por me terem facultado os dados e esclarecido relativamente aos serviços analisados.

A concretização deste trabalho teria sido impossível sem a colaboração de diversas pessoas que tiveram intervenção direta na elaboração do mesmo. No entanto, não posso deixar de referir algumas pessoas que ainda que não tenham sido de alguma forma responsáveis pelo desenvolvimento do projeto, contribuíram bastante na minha formação e nos conhecimentos que fui adquirindo até aqui, razão pela qual tenho de deixar um agradecimento a todos os bons professores que tanto me ensinaram durante o meu percurso académico.
Um agradecimento especial ao fantástico professor Fernando Sequeira, que tive a sorte de ter como professor em algumas cadeiras durante a licenciatura e novamente no mestrado, e com quem posso dizer que aprendi tudo o que sei de Probabilidades.
À minha querida mãe, pela sua compreensão, carinho imensurável e lanchinhos deliciosos durante as tardes de trabalho.
Ao meu super pai, pela motivação e pelos desafios constantes enquanto crescia, que contribuíram para que quisesse fazer mais e melhor.

Muito obrigado a todos pois sem a vossa ajuda não teria sido possível.

Conteúdo

Resumo

Este projeto para a Associação Nacional das Farmácias (ANF) constitui a minha Tese de Mestrado em Matemática Aplicada à Economia e Gestão na Faculdade de Ciências da Universidade de Lisboa.
É um caso real de aplicação de Regressão linear múltipla que tem como principal objetivo identificar as características de cada farmácia que possam estar associadas à prestação de alguns serviços farmacêuticos, ilustrando assim a aplicabilidade dos modelos de regressão linear através de estudos elaborados com base em dados reais.
Os dados utilizados foram fornecidos pela ANF, com informação sobre algumas características das farmácias e de alguns dos serviços prestados pelas mesmas, mantendo no entanto confidencial a identificação da farmácia.
Primeiramente começou-se por fazer uma análise preliminar, organizar e validar a base de dados com a informação relativa às características de cada farmácia e serviços farmacêuticos prestados. Foram utilizados modelos lineares, tabelas de contingência e outras metodologias estatísticas no sentido de perceber a influência de cada característica da farmácia na oferta de alguns serviços mais comuns, bem como no volume de serviços total.

Palavras-chave: Análise de Regressão, Regressão Linear Múltipla, Análise de Variância, Análise de Covariância.

Abstract

This project for the National Association of Pharmacies (ANF) is my master's thesis in Mathematics Applied to Economics and Management at the Faculty of Science, University of Lisbon.
It is a real case of multiple linear regression application that aims to identify the characteristics of each pharmacy that may be associated with some pharmaceutical services, and is also a form of illustrating the applicability of linear regression models through elaborate studies based on real data.
The data were provided by ANF, with information on some characteristics of pharmacies and some of the services provided by them, while maintaining confidential the identity of the pharmacy.
We started by making a preliminary analysis, organize and validate the database. Linear models, contingency tables and other statistical methodologies were used in order to understand the influence of each characteristic in providing some common services as well as the total volume of services .

Keywords: Multiple Linear Regression, Regression Analysis, Analysis of Variance, Covariance Analysis.

Lista de acrónimos e abreviaturas

ANF: Associação Nacional das Farmácias
EMQ: Estimadores dos mínimos quadrados
G.M.: Gauss-Markov
RLS: Regressão linear simples
RLM: Regressão linear múltipla
ANOVA: Análise de variância (ANalysis Of VAriance)
ANCOVA: Análise de covariância (ANalysis Of COVAriance)
i.i.d.: independentes e identicamente distribuídas
IMC: Índice de Massa Corporal

Capítulo 1

Introdução

Ao longo dos últimos vinte anos, as farmácias portuguesas investiram na modernização dos espaços físicos, na formação contínua e em intervenções profissionais estruturadas, acumulando conhecimento e experiência com impacte positivo para a população em iniciativas como a medição da pressão arterial e da glicemia nas farmácias, programas de redução de riscos para a saúde pública como a recolha de seringas usadas, de proteção ambiental e redução de desperdícios como a recolha de medicamentos fora do prazo de validade para incineração, além de inúmeras campanhas de promoção da saúde e prevenção da doença.

Para além da dispensa de medicamentos, a oferta de serviços farmacêuticos nas farmácias portuguesas, de acordo com as necessidades dos doentes, serve para contribuir para o uso seguro do medicamento, para a obtenção do benefício terapêutico pretendido e diminuição do seu desperdício.

Os Serviços Essenciais são serviços de intervenção farmacêutica que englobam todos os serviços prestados por farmacêuticos ou técnicos sob supervisão do farmacêutico, de forma sistemática, durante o ato de dispensa ou atendimento regular. Estes serviços podem distribuir-se segundo dois níveis de intervenção profissional: dispensa de medicamentos e de outros produtos de saúde e, determinação de parâmetros na farmácia e intervenção farmacêutica (Serviços Checksaúde).

O objetivo do primeiro nível de intervenção profissional é assegurar que, no ato da dispensa orientada de medicamentos e/ou produtos de saúde, o doente adquire toda a informação que necessita para utilizar o produto com segurança e obter o benefício terapêutico adequado.

Um serviço Checksaúde consiste na determinação de parâmetros, como por exemplo o IMC, a pressão arterial, glicemia, colesterol total, triglicerídeos entre outros. Estes parâmetros servem de indicador para a intervenção profissional na farmácia, em particular

junto dos doentes no intervalo entre consultas médicas e não substituem as determinações analíticas laboratoriais. A disponibilização de um serviço de determinação de parâmetros mediada por um profissional habilitado, a identificação de indivíduos não medicados com perfil de risco suspeito e a vigilância periódica de doentes (medicados e não medicados) são alguns dos objetivos na oferta destes serviços.

Neste trabalho pretendemos analisar uma base de dados fornecida pela Associação Nacional de Farmácias, com informação acerca de 14 características de uma amostra de 2152 farmácias nacionais (Portugal continental e ilhas). O levantamento dos dados foi efetuado pela equipa de apoio aos associados da ANF, através das visitas regulares que efectuam junto das farmácias associadas.

O número total de farmácias em Portugal é, segundo informação recolhida no final de 2014, igual a 2919, das quais, 2767 são associadas da ANF. Foi contabilizado um total de 2423 farmácias que enviaram informação comercial, isto é, que estabeleceram comunicação informática, com ou sem venda de serviços, sendo que apenas as 2152 farmácias que efectivamente considerámos, têm registo de venda de serviços. Desta forma, a base de dados que vamos analisar representa cerca de 75% da população de farmácias nacionais.

O objetivo principal do estudo é o de perceber a influência de cada característica da farmácia na oferta de alguns serviços mais comuns, bem como no volume de serviços total.

Com o intuito de atingir o objetivo enunciado, começaremos por abordar a teoria da análise de regressão, procurando clarificar este conceito, e aprofundar conteúdos teóricos com interesse para a fundamentação da metodologia adotada no nosso estudo para que, posteriormente façamos a aplicação desta metodologia estatística ao problema em estudo. Por fim, serão apontadas as conclusões mais pertinentes do estudo, bem como algumas sugestões que se considerem adequadas.

Foram considerados nesta análise quatro serviços farmacêuticos: CheckSaúde Colesterol total, CheckSaúde Glicemia, CheckSaúde Pressão arterial e um último serviço, que designámos por "Administração de Injectáveis" e que engloba serviços de administração de vacinas e outros medicamentos.

Relativamente às catorze variáveis categóricas a considerar, seis são nominais e oito são ordinais.

As variáveis categóricas nominais são: Acessibilidade, Instalações, Localização, Meio envolvente, Meio onde se insere e Tipo de utentes.

No que diz respeito a variáveis ordinais temos a Área, Dimensão, Poder de compra dos utentes, Ter ou não gabinete especializado, Proximidade ao centro de saúde, Números de montras, Quadro farmacêutico e Quadro pessoal da farmácia.

Estas catorze características subdividem-se em pelo menos três categorias conforme podemos constatar na tabela imediatamente abaixo.

Acessibilidade à farmácia	Acesso pedonal Fácil estacionamento Interface de transportes públicos Passeios
Área de atendimento ao público	Grande ($> 40\ m^2$) Média (20-40 m^2) Pequena ($< 20\ m^2$)
Dimensões da Farmácia	Grande ($> 100\ m^2$) Média (85-100 m^2) Pequena ($< 85\ m^2$)
Instalações da Farmácia	Antigas; Modernas; Regulares
Localização da Farmácia	Centro Comercial Praça grande / avenida Praceta Rua
Meio envolvente da Farmácia	Escritórios / Serviços Misto Residencial
Meio onde a Farmácia se insere	Rural Semi-urbano Urbano Zona balnear
Número de montras	0; 1; 2; >2
Utentes da Farmácia	De passagem Misto Residentes na área
Poder de compra dos utentes	Baixo; Médio; Elevado
Possui gabinete (sala fechada) de atendimento especializado?	Não Sim (mais de 7 m^2) Sim (menos de 7 m^2)
Proximidade ao Centro de Saúde / Hospital	Muito próxima Relativamente próxima Distante
Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)	1-3 4-6 6-10 > 10
Quadro farmacêutico da Farmácia	1; 2; 3; 4; >4

Capítulo 2

Regressão

2.1 Um pouco de história

A origem do termo "regressão" remonta a Francis Galton (1822-1911), que o empregou pela primeira vez no final do século XIX, num estudo sobre a relação entre a altura dos pais e dos filhos.
Apesar de haver uma tendência para os pais altos terem filhos altos e os pais baixos terem filhos baixos, o cientista inglês observou que filhos de pais muito altos, em média, não eram tão altos quanto os seus pais, e que filhos de pais muito baixos, em média, não eram tão baixos quanto os seus pais. Foi a partir dessas observações que o primo de Charles Darwin concluiu que a altura dos filhos tendia para a média (μ) da espécie e, a cada geração, demonstrou que a altura dos filhos não tende a refletir a altura dos pais, mas sim a regredir para a altura média da população.
A lei de regressão universal de Galton é mais tarde confirmada por um dos seus seguidores, Karl Pearson, quando depois de recolher mais de mil registos das alturas de indivíduos pertencentes a grupos de famílias altas e de famílias baixas, notou que tanto os filhos de pais altos como os de pais baixos "regrediram" em direção à altura média da população e por isso a lei foi chamada de "regressão para a média".
Apesar de no contexto atual, este termo ter muito pouco a ver com essas origens, a denominação permaneceu e por questões históricas o termo é utilizado até hoje.

2.2 Análise de Regressão

Na maior parte dos estudos estatísticos existe a necessidade de estabelecer relações entre as variáveis para que seja possível prever uma ou mais variáveis em termos das outras. Esta necessidade resulta do facto de precisarmos de recolher todas as informações e dados possíveis sobre um determinado fenómeno que temos como objetivo estudar, para que possamos analisá-los e obter conclusões e, na maior parte das situações a recolha de dados de toda uma população não é possível pois além de ser um procedimento moroso é também muito dispendioso.
É assim que surge a Análise de Regressão, que permite, ao ajustar algum tipo de função matemática usando os dados recolhidos, modelar o comportamento dos dados de que não dispomos. Esta poderosa ferramenta é um dos métodos mais populares da análise estatística, com aplicações em praticamente todas as áreas onde a Estatística tem alguma utilidade. Na verdade, mais do que um método estatístico, a análise de regressão pode ser vista como um conjunto de métodos estatísticos que inclui estimação, testes de hipóteses e previsão.
A análise de regressão linear estuda a relação entre uma variável, a variável dependente ou variável resposta e uma ou várias variáveis que designamos por independentes ou explicativas. Esta relação representa-se por meio de um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.
Designamos por modelo de regressão linear simples o modelo onde consideramos uma relação linear entre a variável dependente e uma variável independente. Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se modelo de regressão linear múltipla.

2.3 Modelação estatística

Já vimos que quando temos como objetivo o estudo de um fenómeno, necessitamos de construir um modelo matemático que relacione uma ou mais variáveis resposta a um conjunto de variáveis preditoras. No entanto, devemos encarar a modelação estatística como um processo onde vamos tendo em conta modelos alternativos que vão sendo ajustados, afinados e até mesmo substituídos por outros até encontrarmos o modelo adequado.
Os objetivos de uma modelação podem ser muitos. Predição, estimação e inferência são alguns exemplos.
A Predição é possivelmente a aplicação mais comum dos modelos de regressão. Fazer predição é utilizar o modelo disponível para produzir informação sobre a média da resposta,

para determinados valores das variáveis preditoras, incluindo valores que não fizeram parte do estudo mas pertencem ao intervalo estudado. É assim que são feitos estudos para predizer as vendas futuras de um produto em função do seu preço, a perda de peso de uma pessoa como consequência do número de dias que se submete a uma determinada dieta, a despesa de uma família com médico e remédios em função da renda, a produção de uma determinada cultura em função da quantidade de nutriente aplicada no solo, entre muitos outros. O ideal seria que pudéssemos predizer uma quantidade exatamente em termos de outra, mas isso raramente é possível de modo que na maioria dos casos devemos contentar-nos com a predição de médias, ou valores esperados. (intervalo de previsão)
Em geral, o ajustamento de um modelo de regressão tem por objetivo, além da estimação dos parâmetros, a realização de inferências sobre eles, tais como, testes de hipóteses e intervalos de confiança. Dado um modelo e um conjunto de dados referente às variáveis resposta e preditoras, estimar parâmetros ou ajustar um modelo aos dados significa obter valores ou estimativas para os parâmetros, por algum processo, tendo por base o modelo e os dados observados. Enquanto a predição diz respeito ao uso do modelo como um todo, muitas vezes o interesse pode estar nos valores dos parâmetros em si. É aí que entra a inferência sobre os parâmetros.
Pode também haver um objetivo de melhor compreensão qualitativa do fenómeno que está a ser modelado e/ou da influência que as variáveis preditoras possam ter sobre a variável resposta. Frequentemente, não se tem idéia de quais são as variáveis que afetam significativamente a variável resposta. Para responder a esse tipo de questão é possível conduzir estudos com finalidades exploratórias considerando um conjunto de variáveis e utilizando a análise de regressão como auxilio no processo de seleção das variáveis, eliminando aquelas cuja contribuição não tenha um efeito significativo sobre a resposta.
A selecção de variáveis também pode ser considerado como um objectivo na construção de modelos de regressão. Em muitos casos práticos não fazemos ideia de quais as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão são realizados estudos com um grande número de variáveis e a análise de regressão pode auxiliar no processo de seleção de variáveis, eliminando aquelas cuja contribuição não seja considerada importante.

2.4 Abusos da análise de regressão

É muito importante que tenhamos sempre presente que um modelo ajustado é construído a partir de uma base de dados e, portanto, devemos ter sempre em atenção as limitações inerentes a essa base dados. A análise de regressão é sem dúvida uma pode-

rosa ferramenta estatística mas é por vezes utilizada abusivamente e o mau uso desta ferramenta está normalmente associada à extrapolação, generalização e à determinação de uma relação de causa e efeito.

Uma vez que esperamos que grande parte da variação da variável de saída seja explicada pelas variáveis de entrada, podemos utilizar o modelo para obter valores de Y, variável dependente, correspondentes ao valor de X, variável independente (no caso de RLS), ou valores de X_i, $i = 1, ..., k$, variáveis independentes (no caso de RLM) que não estavam entre os dados. Em geral, usamos valores de X ou X_i que estão dentro do intervalo de variação estudado e já nos referimos anteriormente a esse procedimento que tem o nome de predição. Já a utilização de valores fora desse intervalo tem o nome de extrapolação e deve ser usada com muito cuidado, pois o modelo ajustado pode não ser correto fora do intervalo estudado.
Um conjunto de pontos dá evidência de linearidade apenas para os valores de X cobertos pelo conjunto de dados. Para valores de X que saem fora dos que foram cobertos não há garantia de linearidade. É por este motivo que é arriscado usar uma reta de regressão estimada para predizer valores de Y correspondentes a valores de X que saem fora do âmbito dos dados. (ver figura 1)

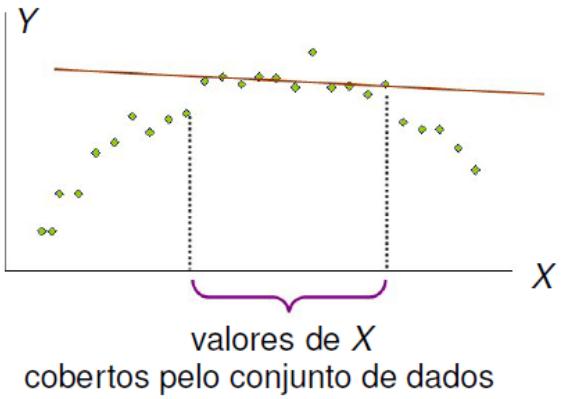


Figura 1

O perigo de extrapolar para fora do intervalo de variação das observações é que a relação existente entre as variáveis pode não se verificar fora do intervalo estudado.
Portanto, extrapolar numa análise de regressão significa utilizar o modelo para predizer ou prever o comportamento da resposta fora da amplitude dos valores das variáveis ex-

plicativas utilizadas no estudo.

A generalização refere-se à extensão das conclusões para populações que não são representadas pela amostra. Devemos sempre fazer uma descrição detalhada da população de estudo e indicar que as conclusões são válidas apenas para essa população. Um aspeto que devemos ter em conta é que bases de dados com poucas observações facilmente conduzirão a conclusões que ultrapassarão os limites adequados.

Outro dos problemas referidos é o de determinar relações de causa e efeito. É muito frequente tentar encontrar relações de causa e efeito por exemplo entre características de um sistema para uma pesquisa científica. Mas essas relações são, em geral, difíceis de estabelecer, o que leva a que a interpretação de resultados da análise nestes termos muitas vezes falha. Um bom ajustamento de um modelo não é sinónimo da existência de uma relação de causa efeito entre variáveis preditoras e a variável resposta. Em muitas situações conseguimos indicadores de que ao variar uma variável a outra também varia, mas nada indica que a variação de uma é a causa da variação da outra.
Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, ..., X_p e uma variável resposta Y, com uma relação de causa e efeito.
A relação entre duas variáveis pode ser de dependência funcional (relação de causa-efeito) de uma em relação à outra, isto é, a magnitude de uma das variáveis (variável resposta) é função, ou é determinada pela magnitude da outra variável (independente). No entanto, existem variáveis que se apresentam correlacionadas mas o que se verifica é uma relação de variação conjunta e não uma relação de dependência. Uma relação causal só pode ser afirmada com base em teoria própria do fenómeno sob estudo, e não com base na relação linear estabelecida estatísticamente.

Capítulo 3

Regressão linear múltipla

A regressão linear múltipla é usada para determinar o valor de uma variável dependente contínua baseando-se na sua relação linear com as variáveis independentes. Possibilita-nos ver o efeito conjunto de várias variáveis X_i na variável dependente Y. Regressão múltipla pode ser vista como uma coleção de técnicas estatísticas para construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo e, pode descrever uma grande variedade de situações de interesse prático. Juntando a este facto as boas propriedades dos estimadores de mínimos quadrados conseguimos facilmente compreender que os modelos lineares são uma ferramenta muito poderosa e, das mais utilizadas na inferência e modelação estatística.

3.1 Modelo teórico e seus pressupostos

Na Regressão Linear Múltipla admite-se que as n observações da variável resposta Y são aleatórias e podem ser modeladas como:

$Y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + ... + b_p x_{p,i} + \epsilon_i,\ i = 1, ..., n$

Trata-se portanto de um sistema de n equações lineares a $p + 1$ incógnitas.
Uma forma alternativa de escrever o modelo de RLM é

$Y_i = \sum_{j=1}^{p} b_j x_{ij} + \epsilon_i$ com $x_{i1} = 1$ para $i = 1, ..., n$

Nesta forma de escrever a variável resposta, o parâmetro b_1 corresponde agora ao termo constante, papel desempenhado pelo b_0 na expressão que usámos anteriormente.

Esta será a maneira como escreveremos o modelo a partir de agora, considerando portanto um sistema de n equações lineares a p incógnitas.
Para encontrar a solução é mais simples reescrever o sistema em notação matricial. As n equações correspondem a uma única equação matricial: $\mathbf{Y} = \mathbf{Xb} + \epsilon$.
Na equação matricial $\mathbf{Y} = \mathbf{Xb} + \epsilon$, designamos por:

- $\mathbf{Y}$ o vector n x 1 das observações da variável dependente; $\mathbf{Y}' = [Y_1 ... Y_n]$
- $\mathbf{X}$ a matriz de planeamento de dimensões n x p em que a primeira coluna é um vetor de 1´s associada à constante aditiva do modelo e as restantes colunas são dadas pelas observações de cada variável preditora; $\mathbf{X} = [X_{ij}]$, $i = 1, ..., n$ $j = 1, ..., p$
- $\mathbf{b}$ o vector de p x 1 de parâmetros do modelo; $\mathbf{b}' = [b_1 ... b_n]$
- ϵ o vector n x 1 dos n erros aleatórios; $\epsilon' = [\epsilon_1 ... \epsilon_n]$

PRESSUPOSTOS DO MODELO

O modelo de regressão linear múltipla (bem como da regressão linear simples) pressupõe a verificação de algumas condições, que apresentamos em seguida.

1. Os erros ϵ_i são variáveis aleatórias de média zero. $(E(\epsilon_i) = 0)$
2. Os erros ϵ_i são variáveis aleatórias de variância constante. $(Var(\epsilon_i) = \sigma^2)$
3. As variáveis aleatórias ϵ_1, ϵ_2,..., ϵ_n são não correlacionadas. $(E(\epsilon_i, \epsilon_j) = 0$, se $i \neq j)$

   Para a construção de intervalos de confiança e testes de hipóteses que abordaremos seguidamente (na secção "Inferência estatística") necessitamos ainda do seguinte pressuposto:

4. Os erros ϵ_i são i.i.d e seguem distribuição normal: $\epsilon_i \cap N(0, \sigma^2)$.

   Idealmente as variáveis explicativas $X_1, X_2, ..., X_p$ devem ser não correlacionadas ou apresentar uma fraca correlação uma vez que, variáveis explicativas muito correlacionadas podem levar a problemas de multicolinearidade, sobre os quais falaremos na seção "Multicolinearidade".

3.1.1 Estimação dos parâmetros

O Método dos Mínimos Quadrados é uma técnica que procura encontrar o melhor ajustamento para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados. A essa diferença damos o nome de resíduos e portanto, através desde método conseguimos estimar os parâmetros minimizando a soma do quadrado dos resíduos da regressão. Representando por SQ soma dos quadrados dos resíduos, o que se pretende é encontrar o mínimo de

$SQ(\mathbf{b}) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} b_j x_{ij})^2$.

De $\mathbf{Y} = \mathbf{Xb} + \epsilon$ vem, $\epsilon = \mathbf{Y} - \mathbf{Xb}$.
Usando notação matricial temos

$$SQ = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'Y} - \mathbf{2bX'Y} + \mathbf{b'(X'X)b}$$

Derivando a soma dos quadrados dos resíduos em ordem a $\mathbf{b}$ e resolvendo a equação $\frac{\partial SQ}{\partial \mathbf{b}} = 0$ vem,

$$\frac{\partial SQ}{\partial \mathbf{b}} = 0 \Leftrightarrow -2\mathbf{X'Y} + 2(\mathbf{X'X})\mathbf{b} = \mathbf{0} \Leftrightarrow \hat{\mathbf{b}} = (\mathbf{X'X})^{-1}(\mathbf{X'Y})$$

Note-se que

$$-2\mathbf{X'Y} + 2(\mathbf{X'X})\mathbf{b} = \mathbf{0} \Leftrightarrow \mathbf{X'}(\mathbf{Y} - \mathbf{Xb}) = \mathbf{0} \Leftrightarrow \mathbf{X'e} = \mathbf{0}$$

ou seja, no modelo linear os resíduos são ortogonais à matriz de planeamento $\mathbf{X}$.
Vamos ver agora que o zero da derivada corresponde, de fato, a um mínimo.

$$SQ = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = (\mathbf{Y} - \mathbf{X\hat{b}} + \mathbf{X\hat{b}} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{X\hat{b}} + \mathbf{X\hat{b}} - \mathbf{Xb})$$
$$= (\mathbf{Y} - \mathbf{X\hat{b}})'(\mathbf{Y} - \mathbf{X\hat{b}}) + \mathbf{2}(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{X'}(\mathbf{Y} - \mathbf{X\hat{b}}) + (\hat{\mathbf{b}} - \mathbf{b})'(\mathbf{X'X})(\hat{\mathbf{b}} - \mathbf{b})$$

Usando o resultado acima demonstrado de que os resíduos são ortogonais à matriz de planeamento verifica-se a igualdade

$$(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{X'}(\mathbf{Y} - \mathbf{X\hat{b}}) = (\mathbf{Y} - \mathbf{X\hat{b}})'\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}) = \mathbf{e'X}(\hat{\mathbf{b}} - \mathbf{b}) = \mathbf{0}$$

e, a soma de quadrados pode ser escrita na forma simplificada

$$SQ = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) + (\hat{\mathbf{b}} - \mathbf{b})'(\mathbf{X}'\mathbf{X})(\hat{\mathbf{b}} - \mathbf{b})$$

Para além do primeiro termo desta soma não depender de $\mathbf{b}$, os termos da soma no lado direito da igualdade são não negativos, pois são somas de quadrados. Assim, o mínimo será atingido no ponto que anular o segundo termo da soma e portanto, será atingido em $\mathbf{b} = \hat{\mathbf{b}}$.
Assim, se a matriz $\mathbf{X}'\mathbf{X}$ for invertível, o vector dos estimadores de mínimos quadrados dos coeficientes de regressão é dado por:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

Chamamos valores ajustados aos $\hat{y_i} = \sum_{j=1}^{p} \hat{b_j} x_{ij}$. Em notação matricial, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$
Os resíduos são calculados através da diferença entre os valores observados e os valores estimados; $e_i = y_i - \hat{y_i}$ para $i = 1, 2, ..., n$.
Representamos por $\mathbf{e}$ o vector dos resíduos, ou seja, o vector das estimativas dos termos de erro que em notação matricial é dado por

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}$$

Antes de passar às seção onde apresentamos as propriedades estatísticas dos EMQ vamos definir duas matrizes que têm um papel importante na dedução de algumas dessas propriedades.
A matriz $\mathbf{H}$ (*hat matrix*) é uma matriz n x n tal que:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Assim, o vetor dos valores ajustados pode ser escrito como função linear dos vetores observados uma vez que:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

De notar, que a matriz $\mathbf{H}$ é simétrica e idempotente, isto é,

$$\mathbf{H}\mathbf{H} = \mathbf{H}^2 = \mathbf{H}$$

A matriz $\mathbf{M}$, também simétrica e idempotente, é definida à custa da matriz $\mathbf{H}$ e da

matriz identidade de ordem n

$\mathbf{M} = \mathbf{I_n} - \mathbf{H}$.

Note que,

$\mathbf{MX} = (\mathbf{I_n} - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$.

e como tal, o vetor dos resíduos pode ser reescrito como função linear dos erros aleatórios,

$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{MY} = \mathbf{M}(\mathbf{Xb} + \epsilon) = \mathbf{M}\epsilon$.

Vamos mostrar que $\sigma^2\mathbf{M}$ é a matriz de covariâncias dos resíduos pois precisaremos de usar este resultado mais à frente.
Usando a relação $\mathbf{e} = \mathbf{M}\epsilon$ e as condições de G.M. vem,

$Cov(\mathbf{e}) = E[\mathbf{ee'}] = E[\mathbf{M}\epsilon\epsilon'\mathbf{M'}] = \mathbf{M}E[\epsilon\epsilon']\mathbf{M} = \mathbf{M}\sigma^2\mathbf{I_n}\mathbf{M} = \sigma^2\mathbf{M}$ *c.q.d.*

3.1.2 Propriedades estatísticas dos EMQ

Nesta secção abordaremos de forma muito breve as boas propriedades estatísticas dos EMQ. Estas propriedades verificam-se sob a validade das condições de Gauss-Markov, que já foram anteriormente enunciadas.

Juntando a propriedade de linearidade do valor médio à primeira condição de G.M. conseguimos garantir que os EMQ são centrados uma vez que,

$E(\hat{\mathbf{b}}) = E[(\mathbf{X'X})^{-1}(\mathbf{X'Y})] = (\mathbf{X'X})^{-1}\mathbf{X'}E(\mathbf{Y}) = (\mathbf{X'X})^{-1}\mathbf{X'Xb} = \mathbf{b}$

e, exigindo a validade da $2^{\underline{o}}$ e $3^{\underline{o}}$ condições de G.M. podemos calcular a matriz de covariâncias dos EMQ,

$Cov(\hat{\mathbf{b}}) = E[(\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})'] =$
$= E[[(\mathbf{X'X})^{-1}(\mathbf{X'Y}) - (\mathbf{X'X})^{-1}\mathbf{X'}E(\mathbf{Y})][(\mathbf{X'X})^{-1}(\mathbf{X'Y}) - (\mathbf{X'X})^{-1}\mathbf{X'}E(\mathbf{Y})]'] =$
$= E[(\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))'(\mathbf{X'X})^{-1}] =$
$= (\mathbf{X'X})^{-1}\mathbf{X'}Cov(\mathbf{Y})\mathbf{X}(\mathbf{X'X})^{-1} =$

$= \sigma^2((\mathbf{X'X})^{-\mathbf{1}})$

Conseguimos garantir a consistência dos EMQ desde que à medida que a dimensão amostral aumenta e no limite se aproxima de infinito, a soma dos elementos da diagonal principal da matriz $(\mathbf{X'X})^{-\mathbf{1}}$ ou seja, o traço da matriz esteja a tender para zero isto é, desde que $tr(\mathbf{X'X})^{-1} \longrightarrow 0$ quando $n \to +\infty$ que é equivalente a escrever $\lim_{n\to\infty} tr(\mathbf{X'X})^{-1} = 0$.

O Teorema Gauss-Markov garante, que o estimador de mínimos quadrados (EQM) é o estimador linear centrado de variância mínima. Assim, os EMQ do modelo de regressão múltipla são estimadores BLUE (Best Linear Unbiased Estimators) ou seja, de entre todos os estimadores lineares centrados são aqueles que possuem variância mínima.
Na verdade, esta propriedade é ainda válida para qualquer combinação linear dos parâmetros, ou seja, para qualquer parâmetro da forma $a'b$ em que $\mathbf{a}$ é um vector px1 de coeficientes. $a' = [a_1 a_2 .. a_p]$
Se estivermos perante um modelo linear

$\mathbf{Y} = \mathbf{Xb} + \epsilon,$

onde se verificam as condições de Gauss-Markov e,

$\hat{\mathbf{b}} = (\mathbf{X'X})^{-\mathbf{1}}(\mathbf{X'Y})$

for o EMQ do vector de parâmetros $\mathbf{b}$ então, podemos verificar facilmente que o estimador $a'\hat{b}$ é aquele que de entre todos os estimadores lineares centrados para $a'b$, tem variância mínima.
Para ver que assim é, consideramos $c'\mathbf{Y}$ um qualquer estimador linear centrado de $a'b$.
Calculando o seu valor médio temos:

$E(c'\mathbf{Y}) = c'E(\mathbf{Y}) = c'\mathbf{X}b$

Como o estimador é centrado, $E(c'\mathbf{Y}) = a'b$
ou seja,
$c'\mathbf{X}b = a'b$
de onde tiramos a igualdade:
$c'\mathbf{X} = a'$.

A variância deste estimador é dada por:

$$Var(c'\mathbf{Y}) = c'Cov(\mathbf{Y})c = c'\sigma^2 I_n c = \sigma^2 c'c$$

e a variância do estimador dos mínimos quadrados é dada por:

$$Var(a'\hat{b}) = a'Cov(\hat{b})a = \sigma^2 a'(\mathbf{X'X})^{-\mathbf{1}}a$$

Fazendo a diferença entre as variâncias dos dois estimadores vem:

$$Var(c'\mathbf{Y}) - Var(a'\hat{b}) = \sigma^2 c'c - \sigma^2 a'(\mathbf{X'X})^{-\mathbf{1}}a = \sigma^2(c'c - a'(\mathbf{X'X})^{-\mathbf{1}}a)$$

Usando a igualdade acima deduzida ($c'X = a'$) temos:

$$Var(c'\mathbf{Y}) - Var(a'\hat{b}) = \sigma^2(c'c - c'\mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X}'c) = \sigma^2 c'(\mathbf{I_n} - \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X}')c = \sigma^2 c'\mathbf{M}c.$$

Ora, $\sigma^2 c'\mathbf{M}c$ é uma quantidade igual ou superior a zero pois sabemos que $\mathbf{M}$ é uma matriz semi-definida positiva uma vez que $\sigma^2\mathbf{M}$ é a matriz de covariâncias dos resíduos. Desta forma, acabámos de mostrar que:

$$Var(c'\mathbf{Y}) \geq Var(a'\hat{b})$$

ou seja, qualquer que seja o estimador linear centrado para $a'b$, a sua variância é sempre maior ou igual que a variância de $a'\hat{b}$ e portanto, o estimador $a'\hat{b}$ é aquele que, de entre todos os estimadores lineares centrados para $a'b$, possui variância mínima.
As propriedades que acabámos de ver são válidas independentemente da distribuição de probabilidade dos erros aleatórios e das observações da variável dependente, até porque, a demonstração entra apenas com cálculos de valores médios e variâncias não sendo imposta nenhuma outra condição no que diz respeito à distribuição de probabilidade.
Isto para chamar a atenção que o método dos mínimos quadrados produz bons estimadores em condições muito gerais.
Ora, sem o pressuposto da normalidade dos erros aleatórios conseguimos fácilmente mostrar que os EMQ são estimadores BLUE, ou seja, dentro da classe dos estimadores lineares são centrados e de variância mínima. Mas, uma coisa é serem os de variância mínima dentro da classe dos estimadores lineares centrados, outra coisa é serem os de variância mínima, sejam lineares ou não.
Se para além das condições de Gauss-Markov, admitirmos a normalidade dos resíduos, ou seja, se admitirmos que os erros aleatórios (ϵ_i's) são variáveis aleatórias independentes e

identicamente distribuidas (i.i.d.) com distribuição normal, $N(0, \sigma^2)$, podemos concluir que os EMQ são estimadores centrados de variância mínima. Não faremos a demonstração deste resultado aqui mas admitindo o pressuposto da normalidade dos ϵ_i's conseguimos ver que a matriz de covariâncias dos EMQ destes parâmetros é igual ao limite inferior de Cramer-Rao, concluindo assim, que de entre todos os estimadores centrados, os EMQ (que neste caso se prova serem iguais ao de máxima verosimilhança) são os que possuem variância mínima.

3.2 Variáveis qualitativas no modelo de regressão

Na análise de regressão, a variável dependente pode ser influênciada tanto por variáveis quantitativas como por variáveis qualitativas.
As variáveis quantitativas são facilmente mensuradas numa escala o que não acontece com as variáveis qualitativas, uma vez que essas variáveis indicam a presença ou a ausência de uma qualidade ou atributo.
Genericamente, a introdução de variáveis qualitativas num modelo de regressão requer a transformação da variável original em variáveis dummy. As variáveis dummy são variáveis dicotómicas, em regra, codificadas com o valor 1 e 0, que representam respectivamente a presença de uma das modalidades e a ausência das restantes. O número de variáveis dummy deve ser igual a $k - 1$, sendo k o número de categorias da variável qualitativa a introduzir no modelo.
A introdução de variáveis qualitativas (dummy) torna o modelo de regressão linear uma ferramenta muito flexível capaz de lidar com muitos problemas encontrados na prática.
No que diz respeito à sua aplicação, este tipo de variável pode ser usada em modelos simples, em que a única variável explicativa é a própria dummy ou em modelos mais complexos, em que uma variável categórica é desdobrada em duas ou mais variáveis dummy.
Os modelos que incluem como variáveis explicativas somente variáveis qualitativas são chamados de modelos de análise de variância (ANOVA). Os modelos de análise de covariância (ANCOVA) são os modelos que, para além de variáveis qualitativas, incluem também variáveis quantitativas.

3.3 Inferência estatística no modelo de regressão

Como já tinhamos mencionado na secção em que falámos dos pressupostos do modelo de regressão, para os testes de hipóteses e intervalos de confiança que abordaremos já de

seguida, vamos supor que os termos de erro seguem uma distribuição normal: $\epsilon_i \cap N(0, \sigma^2)$. Esta condição é necessária pois desta forma conseguimos obter a distribuição de probabilidade de um conjunto de variáveis aleatórias e também estatísticas de forma a que seja possível a construção de intervalos de confiança e testes de hipóteses para os parâmetros do modelo. Procederemos também à construção de intervalos de confiança para a predição da variável $\mathbf{Y}$ correspondente a um conjunto de valores das variáveis independentes não observadas.
A base para a inferência estatística no modelo linear assenta no seguinte teorema:

Teorema
Seja $\mathbf{Y} = \mathbf{Xb} + \epsilon$ um modelo linear em que $\epsilon = [\epsilon_1 \epsilon_2 ... \epsilon_n]'$ é um vector de variáveis aleatórias i.i.d. com distribuição normal $N(0, \sigma^2)$. Então,

1. O EMQ do vetor de parâmetros $\mathbf{b}$ ou seja, $\hat{\mathbf{b}} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'Y}$ tem distribuição multinormal,[1] $\hat{\mathbf{b}} \cap N(\mathbf{b}, \sigma^{\mathbf{2}}(\mathbf{X'X})^{-\mathbf{1}}$.

2. A variável aleatória

   $$\frac{(n-p)S^2}{\sigma^2} = \frac{\mathbf{e'e}}{\sigma^2}$$

   tem distribuição qui-quadrado com $n - p$ graus de liberdade.

3. $\hat{b}$ e S^2 são independentes.

Não faremos aqui a demonstração deste terorema.

No que diz respeito a aplicações práticas, o interesse está nos coeficientes de regressão ou em combinações lineares dos mesmos. Mas, antes de vermos como são construídos intervalos de confiança e testes de hipóteses para estes coeficientes, vamos aproveitar o ponto 2 do teorema acima e deduzir o intervalo de confiança para a variância dos erros, o que por vezes pode ser útil.
Considerando a variável fulcral

$$\frac{(n-p)S^2}{\sigma^2} \cap \chi^2_{n-p}$$

[1] A distribuição multinormal é uma generalização natural da distribuição normal em que as distribuições marginais e condicionais continuam a ter distribuição normal. A função densidade de probabilidade conjunta da multinormal é dada por:

$$f(x_1, ..., x_k) = \frac{e^{[\mathbf{x}-\mu]'\Sigma^{-1}[\mathbf{x}-\mu]}}{(2\pi)^{k/2}|\Sigma|}, \text{ com } \mathbf{x} = (x_1, ..., x_k)$$

o teorema anterior diz-nos que esta tem distribuição qui-quadrado com $n-p$ graus de liberdade. Então, podemos escrever:

$$P(\chi^2_{\alpha/2;n-p} < \frac{(n-p)S^2}{\sigma^2} < \chi^2_{1-\alpha/2;n-p}) = 1-\alpha$$

onde $\chi^2_{\alpha;n-p}$ representa o quantil de probablidade α da distribuição qui-quadrado com $n-p$ graus de liberdade.
Esta equação pode ser reescrita na forma:

$$P(\frac{(n-p)S^2}{\chi^2_{1-\alpha/2;n-p}} < \sigma^2 < \frac{(n-p)S^2}{\chi^2_{\alpha/2;n-p}}) = 1-\alpha$$

de onde tiramos o intervalo a $(1-\alpha)100\%$ de confiança para σ^2

$$(\frac{(n-p)S^2}{\chi^2_{1-\alpha/2;n-p}}, \frac{(n-p)S^2}{\chi^2_{\alpha/2;n-p}})$$

3.3.1 Coeficientes de regressão: testes e intervalos de confiança

Testes de hipóteses individuais para os coeficientes da regressão são fundamentais para se determinar se cada variável explicativa é importante para o modelo de regressão. Muitas vezes o modelo pode ser mais eficaz com a inclusão ou com a exclusão de novas variáveis.
As hipóteses para testar a significância de qualquer coeficiente de regressão individualmente são dadas por: H_0: $b_j = 0$ Vs H_1: $b_j \neq 0$
Se não rejeitarmos H_0 então, podemos retirar x_j do modelo uma vez que, se não rejeitamos a hipótese de $b_j = 0$ podemos concluir que x_j não influencia a variável resposta de forma significativa.
Agora precisamos da estatística de teste para perceber quando é que devemos rejeitar a hipótese nula.
Tendo em conta o ponto 1 do teorema acima, $\hat{b} \cap N(b, \sigma^2(X'X)^{-1})$.
Assim, $Var(\hat{b_j}) = \sigma^2 z_{jj}$, em que z_{jj} é o j-ésimo elemento da diagonal principal de $(X'X)^{-1}$.
De modo que

$$\frac{\hat{b_j} - b_j}{\sqrt{\sigma^2 z_{jj}}} \cap N(0,1)$$

Usando o ponto 2 do teorema conjuntamente com o resultado que diz que o quociente entre uma variável que segue distribuição normal padrão e a raíz quadrada de uma variável com distribuição qui-quadrado dividida pelo respectivo grau de liberdade, resulta numa variável com distribuição t-student, concluímos que:

$$\frac{\frac{\hat{b_j} - b_j}{\sqrt{\sigma^2 z_{jj}}}}{\sqrt{\frac{\frac{(n-p)S^2}{\sigma^2}}{n-p}}} \cap t_{n-p}$$

Desta forma, sob a validade da hipótese nula, chegamos à estatística de teste que não depende de parâmetros desconhecidos e com distribuição conhecida:

$$T = \frac{\frac{\hat{b_j}}{\sqrt{\sigma^2 z_{jj}}}}{\sqrt{\frac{\frac{(n-p)S^2}{\sigma^2}}{n-p}}} = \frac{\hat{b_j}}{S\sqrt{z_{jj}}} \cap t_{n-p}$$

Consequentemente, a região de rejeição do teste bilateral é dada por:

$$\frac{|\hat{b_j}|}{S\sqrt{z_{jj}}} > t_{n-p}^{1-\alpha/2}$$

em que $t_{n-p}^{1-\alpha/2}$ representa o quantil de probabilidade $1-\alpha/2$ da distribuição t de student com $n-p$ graus de liberdade.

Considerando o cálculo do valor-p, dado pela expressão

2 x $P(T_{n-p} > |t_{obs}|)$

e, fixado o nivel de significância α, rejeitamos H_0 se o valor-p for inferior ou igual a α.
Considerando a estatística $\frac{\hat{b_j} - b_j}{S\sqrt{z_{jj}}}$, um intervalo com $(1-\alpha)100\%$ de confiança para os coeficientes da regressão $b_j, j = 0, 1, 2, \ldots, p$, é dado por:

$$[\hat{b_j} - t_{1-\alpha/2;n-p} S\sqrt{z_{jj}}, \hat{b_j} + t_{1-\alpha/2;n-p} S\sqrt{z_{jj}}]$$

Considerando agora o caso mais geral de um parâmetro que é combinação linear dos $b_j, j = 1, ..., p$, vamos trabalhar com parâmetros θ da forma $\theta = a'b$, em que a é um vector

de constantes.
Já vimos que o estimador BLUE para θ é dado pela correspondente combinação linear dos EMQ, $\hat{\theta} = a'\hat{b}$.
Uma vez que estamos a assumir a normalidade dos erros, sabemos que $\hat{\theta}$ é também o estimador de máxima verosimilhança para θ e, para além disso, é de variância mínima.
Calculando o seu valor médio e variância vem:

$E(\hat{\theta}) = E(a'\hat{b}) = a'E(\hat{b}) = a'b = \theta$
e,
$Var(\hat{\theta}) = Var(a'\hat{b}) = a'Cov(\hat{b})a = \sigma^2 a'(X'X)^{-1}a$

Pelo ponto 1 do teorema sabemos que os EMQ têm distribuição normal multivariada e, portanto, qualquer combinação linear destes estimadores tem também distribuição normal. Desta forma, concluímos que, a variável

$$\frac{\hat{\theta} - \theta}{\sigma\sqrt{a'(X'X)^{-1}a}}$$

tem distribuição normal padrão.
Esta variável não é uma variável fulcral para o parâmetro θ porque depende de σ^2, que é desconhecido. Mas, $\hat{\theta}$ e S^2 são independentes e, usando novamente o ponto 2 do teorema e considerando a variável

$$\frac{(n-p)S^2}{\sigma^2}$$

que tem distribuição qui-quadrado com $n-p$ graus de liberdade, temos que:

$$\frac{\dfrac{\hat{\theta} - \theta}{\sigma\sqrt{a'(X'X)^{-1}a}}}{\sqrt{\dfrac{\dfrac{(n-p)S^2}{\sigma^2}}{n-p}}}$$

é o quociente entre uma normal padrão e a raíz de um qui-quadrado a dividir pelo seu número de graus de liberdade e, portanto, tem distribuição t de student com $n-p$ graus de liberdade.
Simplificando a expressão chegamos à variável fulcral studentizada

$$\frac{\hat{\theta} - \theta}{S\sqrt{a'(X'X)^{-1}a}},$$

que nos conduz ao intervalo de $(1-\alpha)100\%$ de confiança para o parâmetro θ

$$[\hat{\theta} - t_{n-p}^{1-\alpha/2} S\sqrt{a'(X'X)^{-1}a}, \hat{\theta} + t_{n-p}^{1-\alpha/2} S\sqrt{a'(X'X)^{-1}a}]$$

em que $t_{n-p}^{1-\alpha/2}$ representa o quantil de probabilidade $1-\alpha/2$ da distribuição t de student com $n-p$ graus de liberdade.
Para construir testes de hipóteses sobre o parâmetro θ podemos usar a mesma estatística e proceder de modo perfeitamente análogo ao que fizemos anteriormente nos testes de hipóteses num só coeficiente de regressão.

Suponhamos agora que, para um certo conjunto de valores não observados das variáveis independentes x, pretendemos estimar o correspondente valor médio da variável dependente $E(Y|X=x)$ e associar-lhe um intervalo de confiança. Neste caso, basta fazer $\theta = x'b$ em todos os resultados anteriores.

3.3.2 Intervalos de predição

Outra situação comum é a previsão intervalar da própria variável aleatória Y condicionada a um certo vector de preditores $x'^* = [x_1^*, ..., x_p^*]$. Este problema difere dos anteriores pois não se insere num contexto paramétrico.
Agora o que pretendemos é algo mais "instável"pois queremos predizer uma variável aleatória $Y|X = x^*$ que vamos designar por y^*. De acordo com o modelo, a variável aleatória y^* é dada por $y^* = x'^*b + \epsilon^*$ em que ϵ^* tem distribuição normal $N(0, \sigma^2)$.
O erro de predição é dado por $\hat{y^*} - y^* = x'^*\hat{b} - x'^*b - \epsilon^* = x'^*(\hat{b}-b) - \epsilon^*$, constituíndo assim uma combinação linear do vector dos EMQ, que tem distribuição normal multivariada, e do termo de erro ϵ^*, que também segue distribuição normal e é independente do vector de estimadores.

$$E(\hat{y^*} - y^*) = E(\hat{y^*}) - E(y^*) = E(x'^*\hat{b}) - E(x'^*b + \epsilon) = x'^*b - x'^*b = 0$$

Desta forma, podemos concluir que o erro de predição tem distribuição normal de valor médio nulo e, variância igual ao EQM de $\hat{y^*}$.

$$EQM(\hat{y^*}) = E[(\hat{y^*} - y^*)^2] = E[(\hat{y^*} - y^*)(\hat{y^*} - y^*)'] = E[(x'^*\hat{b} - x'^*b\epsilon^*)(x'^*\hat{b} - x'^*b + \epsilon^*)'] =$$
$$E[(x'^*(\hat{b}-b) - \epsilon^*)(x'^*(\hat{b}-b) - \epsilon^*)'] = E[(x'^*(\hat{b}-b) - \epsilon^*)((\hat{b}-b)'x^* - \epsilon'^*)] =$$
$$E[(x'^*(\hat{b}-b)(\hat{b}-b)'x^* - x'^*(\hat{b}-b)\epsilon'^* - \epsilon^*(\hat{b}-b)'x^* + \epsilon^*\epsilon'^*] =$$

$E[x'^*(\hat{b}-b)(\hat{b}-b)'x^*] + E(\epsilon^*\epsilon'^*) =$
$\sigma^2[x'^*(XX')^{-1}x^* + 1]$.

Então, pelo mesmo processo de studentização usado anteriormente chegamos à conclusão que a variável

$$\frac{\hat{y^*} - y^*}{S\sqrt{x'^*(XX')^{-1}x^* + 1}}$$

tem distribuição t de student com $n-p$ graus de liberdade, o que nos permite construir um intervalo de predição para y^* dado por

$$[\hat{y^*} - t_{n-p}^{1-\alpha/2} S\sqrt{x'^*(XX')^{-1}x^* + 1}, \hat{y^*} + t_{n-p}^{1-\alpha/2} S\sqrt{x'^*(XX')^{-1}x^* + 1}]$$

3.3.3 Teste F

Numa Regressão Linear Simples ($Y = b_0 + b_1 X$), se $b_1 = 0$, a equação do modelo é apenas $Y = b_0 + \epsilon$. Neste caso, o conhecimento do preditor X em nada contribui para o conhecimento de Y pois o modelo nulo não tira partido da informação dos preditores.
Numa Regressão Linear Múltipla, o modelo nulo corresponde a admitir que todas as variáveis preditoras têm coeficiente nulo e portanto, quando queremos averiguar acerca da utilidade do modelo precisamos de testar a hipótese de que todos os coeficientes de regressão são nulos, (excepto b_0 que corresponde ao termo constante). Devemos assim realizar um teste sobre as seguintes hipóteses:
$H_0 : b_1 = b_2 = ... = b_p = 0$ Vs $H_1 : \exists j = 1, ..., n$ tal que $b_j \neq 0$
Se não rejeitarmos a hipótese nula então é porque não existem motivos para considerar que os coeficientes de regressão não são todos nulos e portanto o modelo é inútil.
Utilizamos a análise de variância, que se baseia na decomposição da soma de quadrados e nos graus de liberdade associados à variável resposta Y, para avaliar a significância do modelo como um todo.
Antes de construirmos o teste para as hipóteses acima comecemos por analisar algumas propriedades importantes.
Denominando,

$SQ_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ - Soma de Quadrados Total

$SQ_R = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ - Soma de Quadrados da Regressão e,

$SQ_e = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ - Soma de Quadrados dos erros,

vamos agora mostrar que podemos decompor a variabilidade total de Y em duas parcelas: uma representando a variabilidade não explicada pelo modelo de regressão (SQ_e) e a outra a variabilidade de Y que o modelo consegue explicar (SQ_R) ou seja, que é válida a equação $SQ_T = SQ_R + SQ_e$.

$$SQ_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 =$$

$$= \sum_{i=1}^{n}((Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2) =$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$$

$$= SQ_R + SQ_e + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Para mostrarmos a igualdade pretendida basta mostrar que $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$
Ora,
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) =$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i - \bar{Y}\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i$$

uma vez que $\mathbf{X'e} = \mathbf{0}$ e portanto, em particular a soma dos resíduos é nula tendo em conta que estamos perante um modelo com termo constante.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i = \sum_{i=1}^{n} e_i\hat{Y}_i = \hat{\mathbf{Y}}'\mathbf{e} = \hat{\mathbf{b}}\mathbf{X}'\mathbf{e} = \mathbf{0}\ c.q.d$$

Voltanto ao teste de hipóteses acima mencionado,

$$H_0 : b_1 = b_2 = ... = b_p = 0 \text{ Vs } H_1 : \exists j = 1, ..., n \text{ tal que } b_j \neq 0$$

o objetivo é construir um teste para esta hipótese e para isso, precisamos de recordar o ponto 2 do teorema que enunciámos no início deste capítulo que nos diz que a soma do quadrado dos resíduos dividida pela variância dos erros tem distribuição qui-quadrado com $n - p$ graus de liberdade.
Sem grandes demonstrações podemos também concluir que, sob a validade de H_0 a soma

dos quadrados dos desvios à média também tem distribuição qui-quadrado mas, com $n-1$ graus de liberdade. Isto porque, sob a validade da hipótese nula, todos os coeficientes de regressão, excepto o que corresponde ao termo contante, são zero o que faz com que tenhamos $y_i = b_0 + \epsilon_i$ ou seja, as observações da variável dependente constituem uma amostra i.i.d. com média b_0 e variância σ^2.
Tendo em conta que podemos escrever a decomposição da soma de quadrados usando a equação,
$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$,

se dividirmos ambos os membos da igualdade acima por σ^2 vem,

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} + \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sigma^2}.$$

A soma de quadrados da regressão é a diferença entre a soma de quadrados total e a soma de quadrados dos resíduos. Ou seja, é dado pela diferença entre duas variáveis independentes com distribuição qui-quadrado e portanto, podemos concluir que a soma de quadrados da regressão também segue distribuição qui-quadrado com $p-1 = (n-1-(n-p))$ graus de liberdade.
Chegamos então à estatistica

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/(p-1)}{\sum_{i=1}^{n} e_i^2/(n-p)}$$

que, sob a validade da hipótese nula, tem distribuição F com $p-1$ graus de liberdade no numerador e $n-p$ no denominador, uma vez que é dada pela quociente entre duas variáveis independentes com distribuição χ^2 dividida pelo número de graus de liberdade. No que diz respeito à região de rejeição, valores grandes da estatística de teste indicam que as variáveis independentes são muito importantes para explicar a variabilidade das observações e por isso, rejeitamos H_0 para valores grandes da estatística F que correspondem a valores-p associados muito pequenos.
No entanto, isso não significa que o modelo esteja bem ajustado e que não possa ser melhorado juntando mais variáveis ou transformando algumas das que já estão incluídas uma vez que a rejeição da hipótese nula não garante um bom ajustamento do modelo.
É bastante comum representar os resultados num quadro-resumo da regressão designado por tabela ANOVA, que representamos em baixo. Isto porque, a análise da informação usada num teste de ajustamento global ajuda-nos a compreender os passos necessários para a construção de um teste F.
De notar que usamos a designação "Média de quadrados" quando nos referimos ao quo-

ciente entre a soma de quadrados e os respectivos graus de liberdade.

Tabela ANOVA

Fonte de variação	Soma de quadrados	Graus de liberdade	Média de quadrados
Regressão	SQ_R	$p-1$	MQ_R
Resídual	SQ_e	$n-p$	MQ_e
Total	SQ_T	$n-1$	$F: MQ_R/MQ_e$

3.3.4 Modelo completo e modelo reduzido

Muitos autores se referem ao princípio da parcimónia quando falam em modelação. Mas de que se trata afinal o princípio da parcimónia?!

O objetivo na modelação é contruir um modelo que descreva adequadamente a relação entre as variáveis mas, que seja o mais simples possível. De um modo geral, pretendemos um modelo parcimonioso, isto é, um modelo que inclua o mínimo possível de variáveis mantendo a qualidade do ajustamento. Isto porque, para além de modelos com muitas variáveis serem pouco práticos e difíceis de interpretar, quanto maior for o número de parâmetros a estimar maior será a variância dos estimadores do modelo de regressão e, por isso, deve-se evitar a inclusão de variáveis que não tenham um peso significativo na explicação da variável dependente.

Adicionar uma variável ao modelo de regressão tem como consequência um aumento na soma dos quadrados da regressão e uma diminuição na soma dos quadrados do erro. No entanto, a adição de variáveis regressoras também aumenta a variância do valor ajustado $\hat{Y}$ e por isso, devemos ter cuidado na selecção das variáveis independentes a incluir no modelo e considerar somente as que realmente explicam a variável resposta.

Quando dispomos de um modelo de RLM com um ajustamento que consideremos adequado, aplicamos este princípio para saber se será possível obter um modelo com menos variáveis preditoras, sem perda significativa na qualidade de ajustamento.

Considerando o modelo de RLM, $Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$, chamamos modelo reduzido a um modelo de regressão linear múltipla que considere apenas algumas das p variáveis preditoras. Se identificarmos o conjunto das variáveis preditoras que pertencem ao modelo reduzido pelo conjunto R, podemos dizer que o modelo completo e o modelo reduzido são idênticos se $b_j = 0$ para qualquer variável x_j cujo índice não pertença a R.

Para avaliar se um dado modelo difere significativamente do modelo reduzido, precisamos de realizar um teste com as seguintes hipóteses:

$H_0 : b_j = 0,\ \forall j \notin R$ Vs $H_1 : \exists j \notin R$ tal que $b_j \neq 0$

Se não rejeitamos H_0 podemos concluir que o modelo reduzido considerado, com menos variáveis preditoras (mais parcimonioso), pode ser utilizado sem afetar a qualidade do ajustamento.
Caso se rejeite H_0, opta-se pelo modelo completo.
Esta situação só envolve coeficientes b_j de variáveis preditoras isto é, o coeficiente b_0 faz sempre parte do modelo reduzido pois não é relevante do ponto de vista da parcimónia uma vez que a sua presença não interfere com a interpretação do modelo.
Uma estatística de teste para a comparação modelo completo/reduzido envolve a comparação das somas de quadrados dos residuos do modelo completo (ao qual nos vamos referir usando o índice C) e do modelo reduzido (referenciado pelo índice R) e, admitindo que passamos do modelo completo para o modelo reduzido quando retiramos q preditores do conjunto C com $p-1$ preditores ou seja, p parâmetros, a estatística de teste é dada por:

$$F = \frac{(SQ_{eR} - SQ_{eC})/q}{SQ_{eC}/(n-p)}.$$

Sob a validade de H_0 isto é, caso $b_j = 0$, para todas as variáveis x_j que não pertençam ao modelo reduzido, a estatística de teste tem distribuição F de Snedcor com q graus de liberdade no numerador e $n-p$ graus de liberdade no denominador.
Caso os modelos completo e reduzido difiram num único preditor, X_j , o teste F é equivalente ao teste T, já referido anteriormente, com as hipóteses $H_0 : b_j = 0$ Vs $H_1 : b_j \neq 0$.
Na prática, com p preditores podemos considerar $2^p - 2$ modelos reduzidos distintos. Se p for um número pequeno, é possível analisar todos os possíveis subconjuntos mas, para p médio ou grande, essa análise completa deixa de ser viável. No entanto, devemos ter o cuidado de não olhar para o ajustamento do modelo completo e, com base nos testes T onde testamos a significância de cada coeficiente b_j , optar pela exclusão de várias variáveis preditoras em simultâneo. Isto porque, quando testamos a significância dos coeficientes b_j partimos do princípio que todas as restantes variáveis pertencem ao modelo. A exclusão de qualquer preditor altera o ajustamento pois altera os valores estimados dos coeficientes e os respectivos erros padrão das variáveis que permanecem no submodelo. Pode até acontecer que um preditor seja dispensável num modelo completo, mas deixe de o ser num modelo reduzido, ou viceversa.

3.4 Validação do modelo

A análise de regressão não termina uma vez estimados os parâmetros de regressão. A qualidade do modelo ajustado deve ser avaliada antes de se concluir algo acerca do grau de influência das variáveis preditoras na variável resposta, ou utilizar o modelo ajustado para fins preditivos.
Antes de abordarmos alguns métodos e análises que nos permitem julgar sobre a adequabilidade de um modelo linear na descrição de um determinado fenómeno vamos falar um pouco de resíduos.
Os resíduos e a soma de quadrados SQ_e têm um papel importante na análise de regressão e fornecem um estimador com boas propriedades para a variância dos erros como aliás já vimos.
O estudo do comportamento dos resíduos dá-nos indicações importantes sobre a qualidade do ajustamento do modelo e sobre a validade das condições de Gauss-Markov.

3.4.1 Análise dos resíduos

Na análise de regressão linear, partimos do pressuposto que os erros são independentes e seguem distribuição normal de valor médio zero e variância constante σ^2, ou seja $\epsilon_i \cap N(0, \sigma^2), i = 1, ..., n$
É necessário validar as suposições do modelo para que os resultados sejam confiáveis já que toda a inferência estatística no modelo de regressão linear se baseia nesses pressupostos e, se houver violação dos mesmos, a utilização do modelo deve ser posta em causa. Desta forma, podemos dizer que uma análise de regressão linear não fica completa sem o estudo dos resíduos.
Análise dos Resíduos é a designação usada para um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos e tem um papel particularmente importante no que respeita a avaliação da qualidade do ajustamento do modelo bem como a verificação das condições de G.M. e de normalidade.
Quando analisamos os resíduos conseguimos ter uma ideia das discrepâncias entre a realidade observada e o modelo e desta forma podemos obter informações muito importantes para encontrar modelos mais adequados e mais precisos.
Como visto anteriormente, o resíduo (e_i) é dado pela diferença entre a variável resposta observada (Y_i) e a variável resposta estimada ($\hat{Y}_i$), isto é,

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{b_1} x_{1i} - ... - \hat{b_p} x_{pi}$$

A ideia básica da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo.
As técnicas utilizadas para verificar as suposições descritas acima podem ser formais, se consistir na aplicação de testes, ou informais como a análise gráfica.
Usando um gráfico residual, as violações dos pressupostos do modelo não são sempre fáceis de detectar e podem ocorrer apesar dos gráficos parecerem bem comportados. É por isso que as técnicas formais são mais indicadas para a tomada de decisão. Na verdade, o ideal é combinar as técnicas disponíveis ou seja, os testes e a representação gráfica.
A condição da normalidade pode ser verificada representando os resíduos em papel de probabilidades normal. Existem dois tipos de gráficos de probabilidade normal: o Normal P-P Plot que representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros e o Normal Q-Q Plot que representa o quantil de probabilidade esperado se a distribuição fosse normal em função dos resíduos. Se os erros possuírem distribuição Normal, todos os pontos dos gráficos devem posicionarem-se mais ou menos sobre a bissetriz dos quadrantes ímpares.

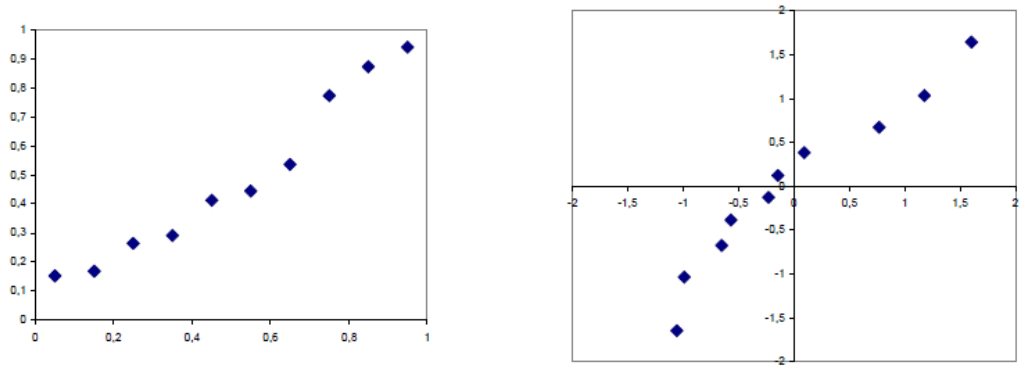


Figura 2

No primeiro gráfico acima (Normal P-P plot) podemos observar que os pontos tendem a concentrar-se em torno da recta de declive 1 que passa na origem, o que dá evidência de que a distribuição dos erros é normal. Da mesma forma, da observação do Q-Q Plot, verifica-se a presunção de normalidade pois os resíduos estão aproximadamente sobre a recta $Y = X$.
Neste tipo de análises é comum usar os resíduos padronizados de forma a terem um desvio

padrão unitário. Mas, é preciso ter em atenção que os novos resíduos não constituem uma amostra aleatória pois não são independentes.
É possível visualizar a forma da distribuição através do histograma e por isso é muito comum comparar o histograma obtido com a função densidade de probabilidade da distribuição normal para perceber se faz sentido assumir a normalidade dos erros.

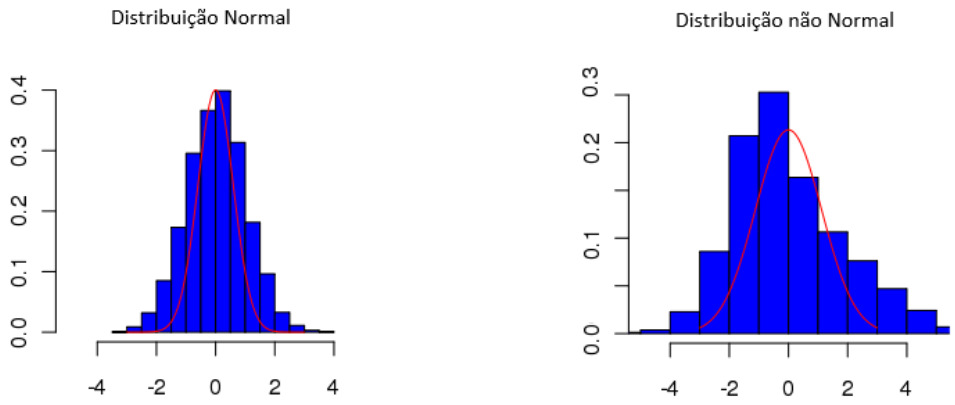


Figura 3

O pressuposto de normalidade pode ainda ser testado analíticamente recorrendo a testes de ajustamento tais como o teste do qui-quadrado, teste de Kolmogorov-Smirnov ou o teste de Normalidade de Lilliefors.
Para a análise formal dos resíduos, podemos ainda realizar o teste de Durbin-Watson para testar independência dos resíduos. Se houver independência, a magnitude de um resíduo não influencia a magnitude do resíduo seguinte e nesse caso, a correlação entre resíduos sucessivos é nula.
Algumas técnicas gráficas para análise dos resíduos consistem na construção de gráficos onde são representados:

a) Resíduos Vs variáveis independentes
Representar e_i Vs x_{ij} para cada j fixo, $j = 1, ..., p$.
Através da análise deste gráfico podemos detectar se alguma das variáveis independentes deverá ser transformada antes de ser incluída no modelo ou se devemos acrescentar ao modelo alguma variável independente que é transformação de uma já existente.
Por exemplo, pode acontecer que a dispersão dos pontos no gráfico se assemelhem à função

exponencial, o que pode sugerir a substituição da variável independente pela exponencial dessa variável.

b) Resíduos Vs outras variáveis independentes não incluídas no modelo.

Este gráfico pode sugerir a existência de algum tipo de relação indicando possivelmente que devemos incluír no modelo a variável em questão.

c) Gráfico dos resíduos Vs valores ajustados

Representar e_i Vs $\hat{Y}_i$

Prova-se que a covariância entre e_i e $\hat{Y}_i$ é nula o que no modelo multinormal equivale a independência. Daí representarmos os e_i Vs $\hat{Y}_i$ e não os e_i Vs Y_i que não são independentes.

Tal como os gráficos que representam os resíduos Vs as variáveis independentes, estes gráficos também ajudam a detectar se existe necessidade de transformar ou juntar variáveis mas, para além disso podemos verificar os pressupostos da independência, média nula e variância constante.

Os pontos do gráfico devem distribuir-se de forma aleatória em torno da recta que corresponde ao resíduo zero, formando uma mancha de largura uniforme. Dessa forma será de esperar que os erros sejam independentes, de média nula e de variância constante. (ver figura 4)

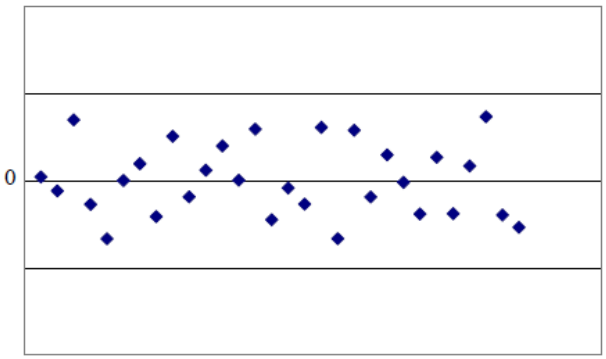


Figura 4

Quando os resíduos não se comportam de forma aleatória, ou seja, seguem um padrão, a condição de independência não é satisfeita.

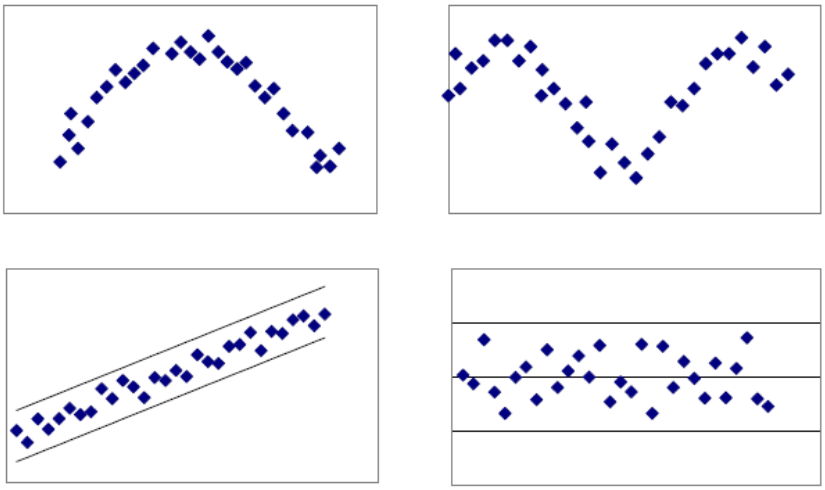


Figura 5

Nos três primeiros gráficos da figura 5 podemos concluir que não existe independência uma vez que os resíduos apresentam comportamentos padronizados.
Já no último gráfico, os resíduos parecem estar distribuídos de forma aleatória, o que sustenta a hipótese da independência dos erros.
A hipótese de homogeneidade de variância dos $e_i's$ deve ser posta em causa se a dispersão dos resíduos aumentar ou diminuir com os valores da variável dependente $\hat{y}_i$.

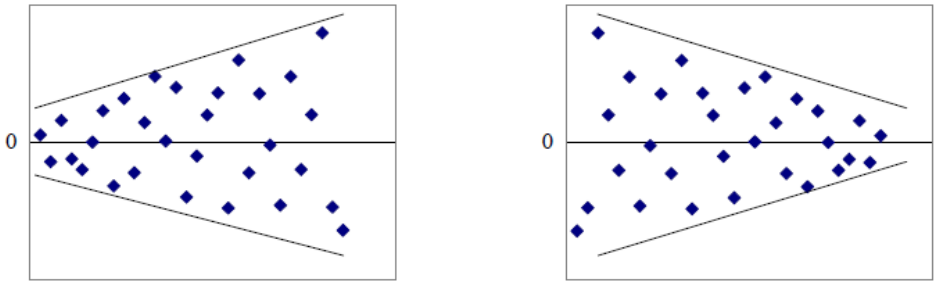


Figura 6

Os gráficos da figura 6 indicam que há violação da hipótese de homogeneidade da variância.
No primeiro gráfico, os resíduos apresentam um comportamento tendencialmente crescente

enquanto que no segundo, o comportamento é tendencialmente decrescente.

3.4.2 O coeficiente de determinação R^2

Tendo em conta a equação acima, segundo a qual podemos decompor a variabilidade total da amostra, introduzimos um coeficiente, que se designa por coeficiente de determinação, e funciona como um indicador do quanto é que a variável resposta é explicada pelo modelo. Uma das formas que temos de avaliar a qualidade do ajustamento do modelo é precisamente atráves deste índice, que assume valores entre 0 e 1. Utilizamos a notação R^2 e definimos o coeficiente de determinação, como a percentagem de variação da amostra que é explicada pelo modelo de regressão. A sua expressão é dada por:

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_e}{SQ_T}$$

Falando um bocadinho dos casos extremos, se a soma dos quadrados dos resíduos for nula, podemos dizer que estamos perante um ajustamento perfeito ($R^2 = 1$),e no caso de SQ_e ser igual a SQ_T, estamos no extremo oposto em que a soma dos quadrados da regressão é nula, e consequentemente $R^2 = 0$. Neste caso, todas as previsões coincidem com a média da variável dependente devido á regressão em nada contribuir para explicar a variabilidade das observações.
Temos de ter em atenção que, o valor do coeficiente de determinação depende do número de observações n e, tende a crescer quando n diminui. Atenda-se ao caso caricato de $n = 2$, em que se tem sempre $R^2 = 1$, o que não significa necessariamente que estejamos perante um ajustamento perfeito.
O R^2 é um índice que deve ser usado com precaução, pois é sempre possível torná-lo maior pela adição de preditores ao modelo e, apesar do R^2 aumentar, isto não significa de todo que o novo modelo seja de facto "melhor"que o anterior. Na verdade, este novo modelo pode inclusivé ser "pior"que o anterior, pois o aumento do R^2 pode ser justificado pela diminuição de 1 grau de liberdade.
A amplitude de variação das variáveis independentes também influência a magnitude deste coeficiente. Em geral, quanto maior a amplitude de variação das variáveis independentes maior o R^2 e quanto menor a amplitude de variação das variáveis independentes menor o R^2.

É possível, olhar para a expressão do R^2 de outra forma e interpretar este coeficiente de maneira diferente. Começemos por reparar que $\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$. Daqui sai,

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Se multiplicarmos por $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, o numerador e denominador deste quociente, otemos

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} =$$

$$= \frac{(\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} = [\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}}]^2$$

Daqui concluimos que, podemos olhar para o coeficiente de determinação como o coeficiente de correlação amostral entre a amostra dos y_i e dos $\hat{y}_i$ e podemos dizer que R^2 mede o grau de associação linear entre os valores observados da amostra e os valores ajustados.

Para evitar dificuldades na interpretação do R^2, alguns estatísticos preferem usar o R_a^2 (R^2 ajustado), que pode ser definido através da seguinte expressão que depende do R^2, da dimensão da amostra n e do número de variáveis independentes p

$$R_a^2 = 1 - (\frac{n-1}{n-p})(1 - R^2)$$

A introdução de mais um preditor no modelo não tem como consequência necessária um aumento do R_a^2 o que acontecia com o R^2.

Esta alternativa pode ser vantajosa quando a dimensão da amostra é pequena, uma vez que já vimos que o valor de R^2 tende a estar inflacionado quando n é muito pequeno. No entanto, o R_a^2 também tem as suas desvantagens. Uma delas é a perda do compromisso entre a soma dos quadrados do erro e a soma dos quadrados da regressão quando o R_a^2 toma valores negativos, o que não acontece com R^2 visto que este coeficiente só toma valores entre 0 e 1.

Se quisermos ser mais rigorosos, devemos considerar o coeficiente de determinação R^2 uma medida da utilidade do modelo, em vez de uma medida da qualidade do ajustamento. Isto porque, se a variância dos termos de erro for grande, o R^2 tende a ser baixo e isso não significa que o modelo esteja mal ajustado mas sim que o modelo é pouco útil, dada a propensão das observações se afastarem de forma significativa do seu valor médio. Pode acontecer também que o modelo incorpore poucos preditores significativos e como consequência tenhamos baixos R^2.

Estamos portanto perante dois casos muito diferentes e é importante que na prática con-

sigamos fazer um diagnóstico do que se está a passar.

Quando estamos perante um modelo sem termo constante, isto é, um modelo em que $b_1 = 0$, a decomposição da soma de quadrados que usámos e a partir da qual definimos o coeficiente de determinação já não é válida e portanto, nesse caso, temos de calcular R^2 de outra forma. Quando estamos perante o modelo sem termo constante, o que sabemos é que
$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n \hat{Y_i}^2 = \sum_{i=1}^n (Y_i - \hat{Y_i})^2 + \sum_{i=1}^n \hat{Y_i}^2$

e portanto, o coeficiente de determinação é definido por

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y_i})^2}{\sum_{i=1}^n Y_i^2} = \frac{\sum_{i=1}^n \hat{Y_i}^2}{\sum_{i=1}^n Y_i^2}$$

3.5 Multicolinearidade

Multicolinearidade é um problema comum na regressão, no qual as variáveis independentes estão relacionadas entre si refletindo redundância de informação nos preditores. A ausência de multicolinearidade é uma das premissas para estabelecer um modelo de regressão múltipla correto. Quando trabalhamos com mais de uma variável regressora, é muito importante verificar se existem relações lineares entre as variáveis explicativas.
Multicolinearidade pode ser explicada como a existência de uma dependência linear forte entre as variáveis independentes e, na sua presença as inferências baseadas no modelo de regressão podem ser pouco confiáveis sendo por isso importante que façamos um diagnóstico de multicolinearidade entre as variáveis de entrada para que a relação existente entre elas não interfira com as estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.
Isto porque, apesar de termos visto que o método dos mínimos quadrados produz estimadores que são ótimos sob as condições de G.M., a qualidade destes estimadores pode ser seriamente afetada se existirem variáveis independentes que estejam linearmente relacionadas entre si e, um mau ajustamento do modelo pode ser uma consequência.
Tomemos como exemplo o modelo:
$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \epsilon_i$,
e suponhamos que existe uma relação linear entre as variáveis preditoras x_{i1} e x_{i2}.
Adimitindo que $x_{i1} \approx x_{i2}$ vem,
$Y_i \approx b_0 + (b_1 + a b_2) x_{i1} + b_3 x_{i3} + \epsilon_i$.
A existência de relações lineares entre as variáveis provoca inflação na variância dos es-

timadores o que indica uma grande instabilidade na inferência afetando desta forma a qualidade do ajustamento.

Em situações em que essas dependências sejam fortes dizemos que existe multicolinearidade. O que precisa de ser feito é procurar variáveis independentes que tenham baixa multicolinearidade com as outras variáveis independentes, mas também apresentem correlações elevadas com a variável dependente.

O indício mais claro da existência da multicolinearidade é quando o R^2 é bastante alto, mas nenhum dos coeficientes da regressão é estatisticamente significativo segundo a estatística T convencional.

O facto das colunas da matriz X serem (quase) linearmente dependentes faz com que possam existir problemas numéricos no cálculo de $(X'X)^{-1}$ (uma vez que o determinante desta matriz é nulo ou um valor muito próximo de zero) e como consequência podemos ter problemas com o ajustamento do modelo e na estimação dos parâmetros.

É possível eliminar o problema da multicolinearidade excluindo da análise uma ou várias variáveis preditoras que sejam responsáveis pela (quase) dependência linear dos preditores. Se a matriz $X'X$ é singular, isto é, algumas variáveis explicativas são combinações lineares de outras, então estamos na presença de multicolinearidade e não há EMQ único para os parâmetros. Se $X'X$ é aproximadamente singular, temos multicolinearidade aproximada. A presença de multicolinearidade pode ser detetada de várias maneiras. Duas medidas que são muitas vezes utilizadas na deteção da multicolinearidade são a tolerância de uma variável e o seu inverso que definimos por fator de inflação da variância.

$TOL_j = 1 - R_j^2$ é a expressão que define a tolerância da variável X_j. R_j representa o valor de R^2 quando se faz a regressão de X_j sobre o conjunto das restantes variáveis.

Assim, se TOL_j estiver próxima da unidade, é porque R_j é um valor próximo de zero e isso significa que a variável X_j é independente das restantes. Por outro lado, se a tolerância assumir um valor próximo de zero, ficamos com a indicação da existência de uma relação aproximadamente linear entre X_j e alguma das outras variáveis independentes.

Utilizamos a designação VIF_j para designar factor de inflacção da variância (variance inflation factor), que não é mais do que o inverso da tolerância. $VIF_j = \dfrac{1}{TOL_j}$.

É uma medida do grau em que cada preditor é explicado pelas restantes variáveis independentes.

No que diz respeito à interpretação, um valor de VIF_j próximo da unidade leva-nos a concluir a ausência de multicolinearidade uma vez que, um VIF próximo da unidade é consequência de TOL_j próximo da unidade que, como já vimos indica que não há dependência entre a variável X_j e as restantes. Valores grandes desta medida indicam que estamos na presença de multicolinearidade. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Geralmente, o VIF é indicativo de problemas

de multicolinearidade se $VIF > 10$ mas alguns autores sugerem que os fatores de inflação da variância não devem exceder 4 ou 5.
Pode-se demonstrar que o VIF também pode ser definido pelo j-ésimo elemento da diagonal principal da matriz R^{-1}. Assim, se optarmos por analisar a matriz R^{-1} devemos retirar do modelo as variáveis correspondentes a entradas muito grandes nesta diagonal. Se houver mais do que um elemento nestas condições, deve-se retirar a variável que corresponde ao maior deles e, em seguida, recalcular a matriz e verificar se existe ainda algum elemento na diagonal demasiado elevado. Se existir, essa variável também deverá ser retirada do modelo e deve-se voltar a calcular a matriz procedendo deste modo até que todas as variáveis que causam multicolinearidade sejam retiradas.

3.6 Seleção de variáveis

Um problema importante na análise de regressão consiste na seleção das variáveis independentes que farão parte do modelo.
Quando as variáveis a incluir no modelo são todas independentes entre si, o processo de seleção de variáveis é muito simples. Podemos por exemplo, ajustar um modelo incluindo todas as variáveis e eliminar aquelas cujo teste t seja significativo. Mas, quando a multicolinearidade é muito forte, todos os métodos podem funcionar mal e conduzir a modelos que não são os melhores. Assim, a seleção de variáveis na presença de multicolinearidade pode ser um problema, mesmo quando esta não é muito forte e, em muitos problemas é quase inevitável a existência de algum relacionamento entre variáveis independentes.
Os métodos de seleção de variáveis são muito úteis nesses casos. Existem alguns métodos distintos na seleção dos preditores no modelo de regressão. Não se pode dizer que haja um método que produza melhores resultados do que os outros e, em princípio, é conveniente utilizar mais de um método e escolher aquele que faz mais sentido ou que apresenta um melhor ajustamento. Por vezes, métodos diferentes conduzem a resultados diferentes mas semelhantes e dificilmente se pode dizer qual dos modelos é melhor. Outras vezes, diferentes métodos produzem o mesmo modelo, o que é uma boa indicação da qualidade do ajustamento.
Os métodos de seleção de variáveis mais utilizados são aqueles que vão procurando, uma a uma e passo a passo, as variáveis a introduzir no modelo ou a eliminar, analisando os efeitos de cada decisão. Apresentamos em seguida, de forma resumida, três métodos: Forward, Stepwise e Backward.

MÉTODO DA SELEÇÃO PROGRESSIVA ("Forward Selection")

O método da seleção progressiva começa com 0 variáveis no modelo e vai incluindo, sucessivamente, aquelas que provocam um maior aumento na qualidade do ajustamento.
O procedimento deste método de seleção baseia-se no princípio de que os preditores devem ser adicionados ao modelo um de cada vez até que mais nenhum seja considerado significativo.

MÉTODO DA SELEÇÃO "STEPWISE"

O método Stepwise é provavelmente a técnica mais utilizada de seleção de variáveis. A regressão Stepwise começa por construir um modelo com uma variável, usando a variável independente que está mais correlacionada com a variável resposta. Em cada etapa deste procedimento são construídos iterativamente uma sequência de modelos de regressão, adicionando ou removendo variáveis. O critério para adicionar ou remover uma variável em qualquer etapa é geralmente expresso em termos de um teste parcial F.
O método de seleção stepwise pode ser visto como um método de seleção progressiva ao qual se junta, após a inclusão de uma nova variável, um novo passo em que se testa a significância de todas as variáveis incluídas no modelo e se retiram aquelas que não forem significativas.

MÉTODO DA SELEÇÃO REGRESSIVA ("Backward Selection")

Enquanto o método Forward começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método Backward faz o oposto. Começa por incorporar inicialmente todas as variáveis e depois, cada uma pode ou não ser eliminada em cada etapa do procedimento. A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo.
O método da seleção regressiva é portanto um método simples que começa por incluir todas as variáveis no modelo e elimina, progressivamente, aquelas cujo teste t é menos significativo, ou seja, aquelas variáveis cuja influência no ajustamento do modelo é a menor. Depois de eliminar a variável menos significativa, o modelo vai sendo reajustado com as restantes variáveis, repetindo-se o processo até que todas as variáveis sejam consideradas significativas.
Naturalmente que, em qualquer dos métodos de seleção utilizados, o modelo final depende do nível de significância escolhido para o teste t, mas é usual considerar α=0.05.

Capítulo 4

ANOVA e ANCOVA

4.1 Análise de variância (ANOVA)

A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas.
Mas, uma variável resposta numérica pode depender de uma ou mais variáveis qualitativas (categóricas), ou seja, de um ou mais fatores. Em tais situações pode ser útil uma Análise de Variância (ANOVA).
Existem situações em que interessa comparar os efeitos médios de mais do que dois tratamentos. Supondo que dispomos de k amostras provenientes de populações gaussianas e homocedásticas, a tentação natural é usar o teste T e proceder à comparação dos pares dos efeitos médios.
É no entanto uma metodologia errada porque torna mais provável tomar decisões erradas. Temos que ter em linha de conta que as decisões estatísticas são decisões de risco, e a metodologia dos testes de hipóteses assenta na construção de regiões de rejeição que fixam a probabilidade α do erro de primeira espécie. Se procedermos a comparações múltiplas, rapidamente o nível do teste se aproxima de 1 e à medida que o número de combinações dois a dois aumenta torna-se quase certo rejeitar alguma hipótese nula verdadeira.
Em outras palavras, há um risco α de rejeitar a igualdade dos efeitos médios de dois tratamentos, quando de facto não o deveríamos fazer. Mesmo que α assuma um valor muito próximo de zero, ao fazer a comparação de n tratamentos dois a dois a probabilidade de erro aumenta rapidamente.
Tomemos como exemplo um caso em que pretendamos comparar o efeito médio de 10 tratamentos. Nesse caso, existem $\binom{10}{2} = 45$ comparações, e se o risco de erro assumido em cada uma delas for 0.05, a probabilidade global de erro é cerca de 0.90 $[1-(1-0.05)^{45}]$. Se em vez de 10 considerarmos 20 tratamentos, há $\binom{20}{2} = 190$ pares, e a probabilidade de

erro é cerca de 0.9999. $[1-(1-0.05)^{190}]$
Fisher teve a ideia genial de proceder a todas as comparações simultaneamente, isto é considerar que a hipótese a que deveria condicionar era a igualdade das médias de todos os efeitos. É assim que surge a análise de variância, uma metodologia estatística desenvolvida por este matemático, nos anos 20, que permite avaliar afirmações sobre as médias de populações.
Estudando o quociente de dois estimadores independentes da variância, Fisher conseguiu os resultados que permitem a comparação simultânea dos efeitos médios de mais do que dois tratamentos. Esta área da estatística tem o nome de análise da variância pois assenta na comparação dos estimadores da variância. Tem como principal objetivo verificar se existe uma diferença significativa entre as médias podendo desta forma concluir se os fatores influênciam de alguma forma a variável dependente.
Em outras palavras, esta ferramenta é utilizada quando se quer decidir se as diferenças amostrais observadas são reais isto é, se são causadas por diferenças significativas nas populações observadas ou se serão apenas casuais e decorrem da mera variabilidade amostral. Esta análise parte do pressuposto que o acaso produz apenas pequenos desvios, sendo que as grandes diferenças se devem a causas reais.
É provavelmente a mais usada e, segundo Gilbert (1989), a mais mal usada metodologia estatística.

4.2 Análise de variância como um modelo de regressão

A ANOVA permite fazer a comparação global de diversas amostras ou subamostras minimizando a probabilidade de erro amostral, já que, conforme vimos no exemplo dado na secção anterior, o total de comparações entre pares aumenta exponencialmente com o aumento do número de amostras.
A aplicação da análise de variância pressupõe a verificação das seguintes condições:

1. As amostras devem ser aleatórias e independentes.
2. As amostras devem ser extraídas de populações normais.
3. As populações devem ter variâncias iguais .

O objetivo é avaliar se várias médias populacionais são iguais.
Se considerarmos k amostras independentes provenientes de uma população X_i com distribuição $N(\mu_i, \sigma^2)$ onde cada observação se pode escrever na forma
$X_{ij} = \mu_i + \epsilon_{ij}, i = 1, ..., k$ e, $j = 1, ..., n_i$

com $\epsilon_{ij} \cap N(0, \sigma^2)$ variáveis aleatórias i.i.d., podemos usar uma estatística de teste com distribuição F de Snedcor para o teste de hipóteses:

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ Vs H_1: existe pelo menos um μ_i que é diferente dos outros.

A dedução da estatística de teste será feita mais adiante, considerando que este modelo também pode ser escrito como um modelo de regresão linear múltipla.
De fato, usando notação matricial podemos escrever $Y = Xb + \epsilon$ em que o vetor das observações Y é tal que,
$Y' = [Y_1 Y_2 ... Y_N] = [X_{11} ... X_{1n_1} | X_{21} ... X_{2n_2} | ... | X_{k1} ... X_{kn_k}]$;
a matriz de planeamento X é dada por,

$$X = \left[\begin{array}{cccc} 1 & 0 & ... & 0 \\ 1 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 1 & 0 & ... & 0 \\ \hline 0 & 1 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 1 & ... & 0 \\ \hline ... & ... & ... & ... \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ \hline 0 & 0 & ... & 1 \\ 0 & 0 & ... & 1 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{array}\right]$$

uma vez que é construída através das variáveis indicatrizes x_{ij}, $i = 1, ..., N$ e $j = 1, ..., k$ que assumem o valor 1 ou 0 respectivamente se a observação Y_i pertence ou não á população j; o vector dos coeficientes de regressão b será aqui representado por μ uma vez que é constituído pelos valores médios das I populações, $\mu' = [\mu_1 \mu_2 ... \mu_k]$. As matrizes $X'X$ e $X'Y$ dadas por:

$$X'X = \left[\begin{array}{cccc} n_1 & 0 & ... & 0 \\ 0 & n_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & n_k \end{array}\right]$$

$$X'Y = \begin{bmatrix} \sum_{j=1}^{n_1} X_{1j} \\ \sum_{j=1}^{n_2} X_{2j} \\ \cdots \\ \sum_{j=1}^{n_k} X_{kj} \end{bmatrix}$$

e portanto, os EMQ do vetor dos coeficientes μ são:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu_1} \\ \hat{\mu_2} \\ \cdots \\ \hat{\mu_k} \end{bmatrix}$$

Ora, $\hat{\mu} = (X'X)^{-1}X'Y$ com a matriz $(X'X)^{-1}X'Y$ dada por,

$$(X'X)^{-1}X'Y = \begin{bmatrix} \overline{X_{1.}} \\ \overline{X_{2.}} \\ \cdots \\ \overline{X_{k.}} \end{bmatrix}$$

onde,

$$\overline{X_{i.}} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

representa a média da amostra i e,

$$\overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{N}, \text{ com } N = \sum_{i=1}^{k} n_i$$

designa a média de todas as observações.

Como $X_{ij} = \mu_i + \epsilon_{ij}$, concluímos que, para este modelo, os valores ajustados são dados pela expressão $\hat{X_{ij}} = \hat{\mu_i} = \bar{X_{i.}}$
Nestas circunstâncias, a fórmula da decomposição da soma dos quadrados, sem termos necessidade de fazer mais cálculos, vem:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2 = \sum_{i=1}^{k} n_i(\overline{X_i} - \overline{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X_i})^2$$

de onde tiramos de imediato a igualdade $SQ_T = SQ_F + SQ_e$.

Na análise de variância, as médias das k amostras aleatórias extraídas das k populações em estudo permitem definir dois tipos de variação: a variação entre as amostras SQ_F (também designada por variação entre grupos, entre tratamentos ou factorial), que resulta da influência do factor sobre a variável em estudo, e a variação dentro das amostras SQ_e (também designada variação residual), que resulta da influência de outros factores não controlados.
A quantificação da variação total, factorial e residual é feita através da média dos quadrados que como já vimos se obtém através do quociente entre as somas dos quadrados e os respectivos graus de liberdade. E portanto, é a razão entre as duas fontes de variação (factorial e residual) que permite concluir acerca da igualdade das k médias. Ou seja, a estatística de teste é dada por:

$$F = \frac{MSQ_F}{MSQe} = \frac{SQ_F/k-1}{SQ_e/N-k}$$

Na prática podemos estar perante duas situações.

1. A variação factorial é muito superior à variação residual, ou seja, o quociente entre as duas fontes de variação é elevado, de onde tiramos que as diferenças entre as amostras são significativas e portanto o mesmo deve acontecer com as populações. Nesse caso, a estatística de teste F assumirá um valor significativamente maior que 1 e H_0 será rejeitada.

2. No caso em que H_0 é verdadeira, as diferenças observadas entre as médias amostrais são atribuídas a flutuações amostrais, o quociente entre as duas fontes de variação é baixo e portanto não há razões para acreditar que haja diferenças significativas entre a média as populações.

F tem distribuição F de Snedcor com $k-1$ graus de liberdade no numerador e $N-k$ graus de liberdade no denominador e a região de rejeição deste teste é sempre unilateral, uma vez que rejeitamos H_0 para valores grandes da estatística de teste ou seja, para valores da estatística de teste maiores ou igual a $f_{1-\alpha;(k-1,N-k)}$ que designa o quantil de probabilidade $1-\alpha$ de uma $F(k-1, N-k)$.

O modelo ANOVA é muitas vezes representado através de um outro conjunto de parâmetros quando se considera a reparametrização:

$X_{ij} = \mu + \alpha_i + \epsilon_{ij},\ i = 1, ..., k$ e, $j = 1, ..., n_i$

com,

$$\mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{N} \text{ e, } \alpha_i = \mu_i - \mu, i = 1, ..., k$$

onde é válida a relação

$$\sum_{i=1}^{k} n_i \alpha_i = \sum_{i=1}^{k} n_i \mu_i - N\mu = 0$$

No que diz respeito à terminologia, X_{ij} representa a j-ésima observação medida no tratamento i; μ é a média geral de todas as observações; α_i o efeito do tratamento i e, ϵ_{ij} o erro aleatório.
Com esta reparametrização, os novos parâmetros são função linear dos anteriores, o que faz com que os EMQ são a mesma função linear dos EMQ de $\mu_1, \mu_2, ..., \mu_k,$, ou seja,

$$\hat{\mu} = \frac{\sum_{i=1}^{k} n_i \overline{X_{i.}}}{N} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}}{N} = \overline{X}$$

e,

$$\hat{\alpha_i} = \hat{\mu_i} - \hat{\mu} = \overline{Y_i} - \overline{Y}$$

Apesar desta reparametrização do modelo envolver $k + 1$ parâmetros, $(\mu, \alpha_1, \alpha_2, ..., , \alpha_k)$ só k é que são independentes uma vez que, como $\sum_{i=1}^{k} n_i \alpha_i = 0$ podemos escrever um dos α_i's como combinação linear dos outros e portanto o vector de coeficientes inclui apenas os parâmetros $\mu, \alpha_1, \alpha_2, ..., , \alpha_{k-1}$.
Com esta reparametrização o teste de igualdade de médias pode ser feito considerando as seguintes hipóteses:
$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$ Vs $H_1 : \exists \alpha_i \neq 0$, para algum i.
No que diz respeito à estatística de teste, em ambos os modelos, temos $\hat{Y_{ij}} = \overline{Y_i}$ e portanto, a partição da soma de quadrados é a mesma e, consequentemente, chegamos à mesma estatística F.

Até agora referimo-nos sempre à análise de variância simples (com um fator). Supondo que existem agora dois fatores (A e B) a influênciar uma certa característica e admitindo que existem I níveis diferentes do fator A e J níveis do fator B, podemos escrever o modelo (adaptando a reparametrização acima) da seguinte forma:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \ i = 1, ..., I, \ j = 1, ..., J \text{ e, } k = 1, ..., n$$

onde,

X_{ijk} representa a k-ésima observação medida no tratamento i do fator A e tratamento j do fator B;
μ é a média global de todas as observações;
α_i o efeito do tratamento i;
β_j o efeito do tratamento j;
γ_{ij} a interação entre o fator A e o fator B;
ϵ_{ijk} o erro aleatório.
Quando os valores de γ_{ij} são nulos significa que estamos perante um modelo sem interação que habiualmente se designa por modelo aditivo.
Para este modelo, é válida a relação $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$.

A primeira hipótese a testar é a de existência de interação e portanto começamos por realizar o teste com as hipóteses

$H_0 : \gamma_{ij} = 0$ Vs $H_1 : \gamma_{ij} \neq 0$, $i = 1, ..., I$, $j = 1, ..., J$.

A estatística de teste a utilizar será a estatística F deduzida para modelo completo/reduzido em que consideramos como modelo reduzido o modelo sem interação e, modelo completo o modelo com interação.
Quando rejeitamos a hipótese nula admitimos que não existe interação entre os fatores e portanto, como já foi dito, estamos perante o modelo aditivo e as hipóteses a testar em seguida são:

$H_0^* : \alpha_i = 0$, $i = 1, ..., I$ e, $H_0' : \beta_i = 0$, $j = 1, ..., J$

para descobrir se os fatores A e B influenciam ou não a característica em estudo.

Não faremos aqui a demonstração mas nesta situação (modelo aditivo com 2 fatores) é possível decompor a soma de quadrados total de acordo com a expressão:

$SQ_T = SQ_A + SQ_B + SQ_e$

e chegar à estatística de teste para a hipótese $H_0^* : \alpha_i = 0$,

$$F = \frac{SQ_A/(I-1)}{SQ_e/(nIJ-I-J+1)}$$

De forma análoga chegamos à estatística,

$$F = \frac{SQ_B/(J-1)}{SQ_e/(nIJ-I-J+1)}$$

para testar a hipótese $H_0' : \beta_i = 0$.

Por outro lado, se não rejeitarmos a hipótese de interação entre os fatores, o que é sugerido por muitos autores é reduzir o problema a uma sequência de testes de análise de variância simples e testar o efeito de um fator separadamente para cada nível do outro fator. Ou seja, devemos testar as hipóteses:

$H_0(j) : \mu_{1j} = ... = \mu_{Ij}$, para $j = 1, ..., J$

e,

$H_0(i) : \mu_{i1} = ... = \mu_{iJ}$, para $i = 1, ..., I$.

4.3 Análise de covariância (ANCOVA)

Além da variável dependente, pode existir uma ou mais variáveis quantitativas numa situação de análise de variância.
Essas variáveis podem ser incluídas no modelo como variáveis independentes e, se afetarem de certa forma os resultados são conhecidas como covariáveis.
A Análise de covariância pode ser vista como uma mistura de análise de variância e regressão, sendo aliás muitas vezes descrita desta forma. Esta análise pode ser entendida como um elo de ligação entre a análise de variância e a análise de regressão uma vez que um dos seus principais objetivos é avaliar o efeito de um ou mais fatores explicativos de natureza nominal numa dada variável resposta uma vez removida a influência que um ou mais fatores quantitativos podem também exercer nessa variável. No entanto, demonstra-se que através da especificação de um modelo de regressão com variáveis dummy é possível atingir de uma forma muito eficiente este objetivo particular da análise de covariância.
Regressão Linear e Análise de Variância são casos particulares do Modelo Linear, que inclui também a Análise de Covariância. Em qualquer destas três situações o que se procura é modelar uma variável resposta quantitativa (numérica) Y. É a natureza das variáveis

explicativas que distingue as três situações.

Numa Regressão Linear, os preditores são variáveis igualmente quantitativas (numéricas). Numa Análise de Variância, as variáveis independentes são fatores (variáveis qualitativas, ou categóricas). E, numa Análise de Covariância, encontramos entre as variáveis explicativas, quer variáveis numéricas, quer fatores.

A inclusão destas variáveis independentes no modelo resulta na redução da variância do erro e consequentemente no aumento da precisão, sendo esse o principal motivo para o uso de covariáveis.

A análise de covariância é uma técnica estatística com alguma complexidade pela quantidade significativa de cálculos que envolve. O nível de sofisticação deste método aumenta com o aumento do número de fatores explicativos nominais, cuja significância se pretende testar, bem como com o aumento do número de variáveis quantitativas que importa controlar. Este problema é eliminado quando a experiência em causa pode ser efectuada através de programas informáticos adequados.

A inferência estatística num modelo ANCOVA (estimação, testes, etc) é feita como no capítulo 3 uma vez que, como já foi dito, se trata de de um modelo linear. A questão principal é a definição correta da matriz de planeamento devendo esta ter em conta as possíveis interações entre as variáveis qualitativas bem como as interações entre estas e as variáveis quantitativas.

Capítulo 5

Aplicação ao problema em estudo

5.1 Análise descritiva dos dados

Neste capítulo faremos uma pequena análise descritiva de cada um dos quatro serviços considerados, através do cálculo de algumas medidas de localização e dispersão e representações gráficas.

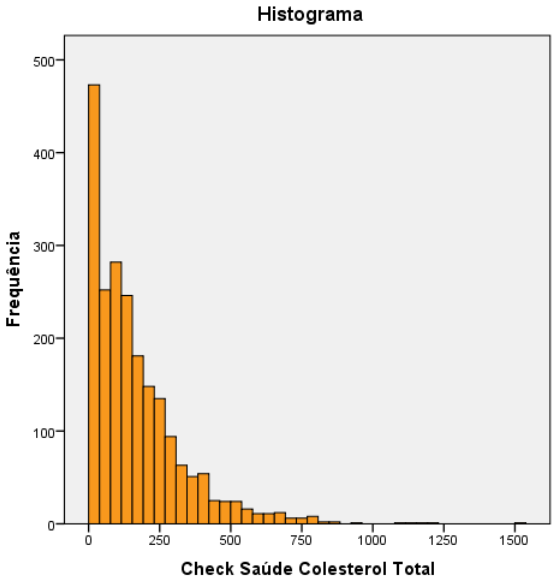
5.1.1 Checksaúde Colesterol Total

O serviço Checksaúde consiste na determinação de parâmetros e existem alguns passos associados à oferta destes serviços.
Em primeiro lugar, a situação do individuo é analisada, com o objetivo de verificar a existência de sinais, fatores de risco ou possíveis queixas.
Depois da determinação dos parâmetros, que é feita com base em procedimentos que visam a execução correta da técnica de determinação e a verificação da exatidão e precisão das determinações efetuadas, os resultados são interpretados com base nos objetivos definidos pelo médico ou, na sua ausência, são usados valores de referência das "guidelines" nacionais ou internacionais que constam dos procedimentos de intervenção profissional.

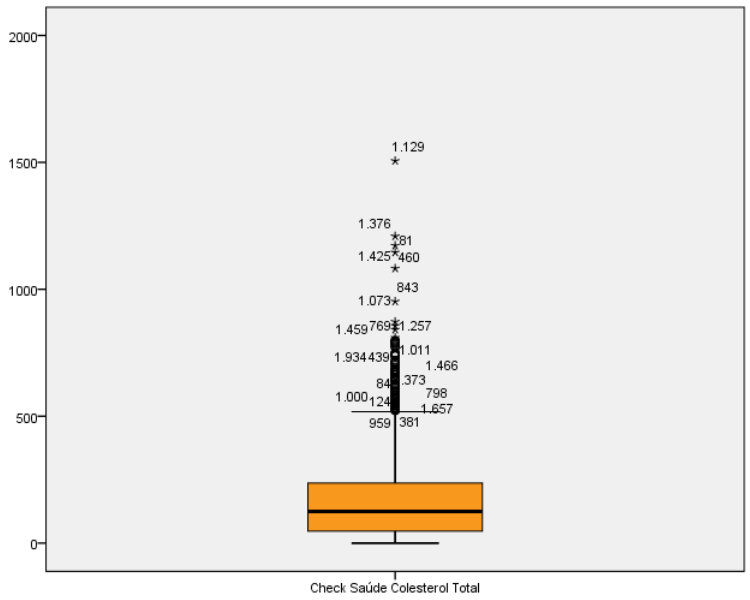
No que diz repeito ao serviço Checksaúde Colesterol Total começemos por analisar algumas medidas de estatística descritiva.

			Estatística	Erro Padrão
Check Saúde Colesterol Total	Média		166,36	3,564
	95% Intervalo de Confiança para Média	Limite inferior	159,37	
		Limite superior	173,35	
	Mediana		125,00	
	Variância		27088,327	
	Desvio Padrão		164,585	
	Mínimo		0	
	Máximo		1506	
	Intervalo		1506	
	Intervalo interquartil		190	
	Assimetria		1,921	,053
	Curtose		6,045	,106

A estimativa para o volume médio de vendas registadas do serviço Checksaúde Colesterol Total é de cerca de 166 unidades, sendo o I.C. a 95% de]159.37;173.35[.
No entanto, existem muitas farmácias sem vendas (inclusive zero é a moda) e alguns pequenos aglomerados periféricos na cauda superior com grande volume de vendas o que implica uma grande variabilidade de comportamento no universo das farmácias ($S' \approx \bar{X}$).
A dispersão é muito elevada em grande parte devido à existência de um grupo de outliers superiores severos, o que faz com que a média seja um valor pouco representativo devido à sua fraca resistência.
O formato do histograma faz lembrar o gráfico de uma exponencial com grande concentração das farmácias na classes com pequeno volume de vendas. Note-se que na distribuição exponencial o valor médio é igual ao desvio-padrão e os valores empíricos obtidos verificam esse facto.



Os "outliers" superiores são responsáveis pelo deslocamento da média para a direita, existindo assim uma assimetria marcada que conseguimos vizualizar quer através do histograma quer através da caixa de bigodes, onde a amplitude do lado esquerdo da caixa é mais pequena que a amplitude do lado direito.

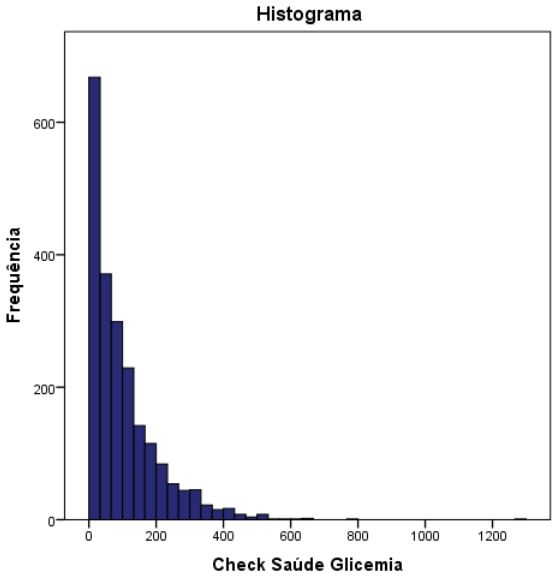


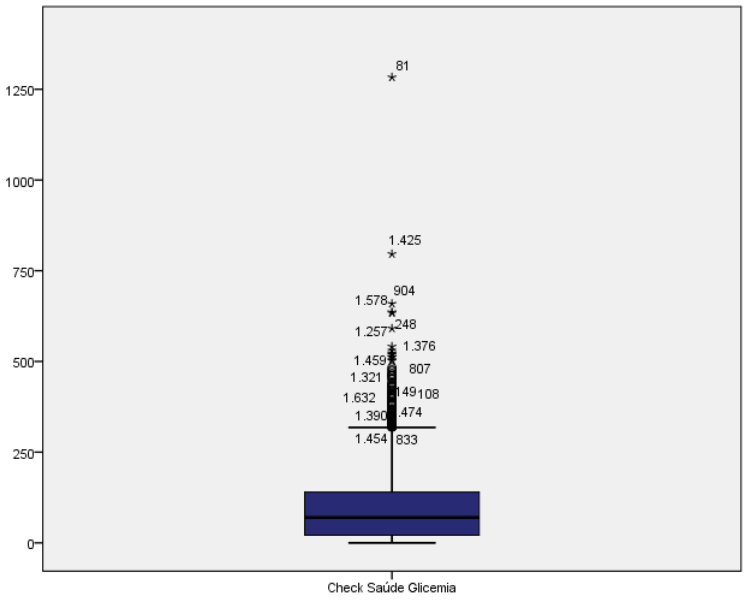
5.1.2 Checksaúde Glicemia

A análise descritiva dos restantes serviços é muito semelhante à do serviço Checksaúde Colesterol Total.

Também no serviço Checksaúde Glicemia se verifica a assimetria positiva marcada, a existência de muitos "outliers" superiores, uma grande variabilidade da variável resposta, e a forma da distribuição sugere-nos novamente a densidade da exponencial.

			Estatística	Erro Padrão
Check Saúde Glicemia	Média		99,11	2,300
	95% Intervalo de Confiança para Média	Limite inferior	94,60	
		Limite superior	103,62	
	Mediana		70,00	
	Variância		11278,816	
	Desvio Padrão		106,202	
	Mínimo		0	
	Máximo		1283	
	Intervalo		1283	
	Intervalo interquartil		119	
	Assimetria		2,202	,053
	Curtose		10,322	,106

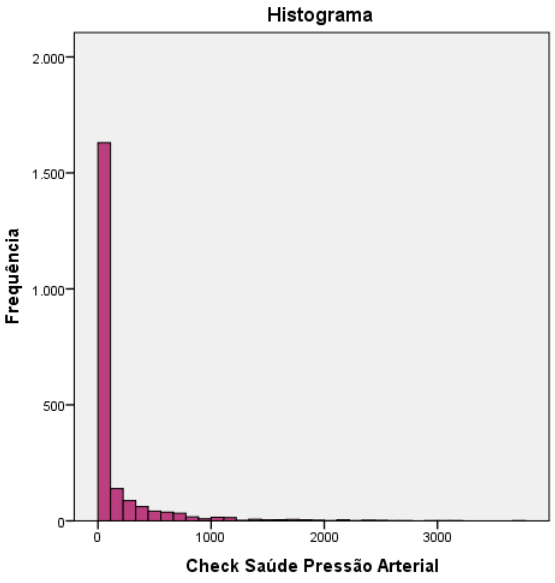


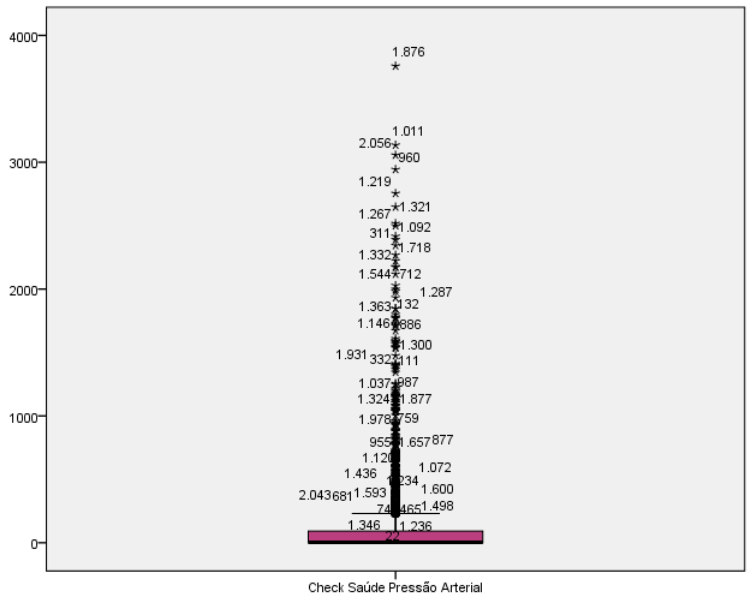


5.1.3 Checksaúde Pressão Arterial

No serviço Checksaúde Pressão Arterial, a assimetria é bastante mais acentuada que nos serviços anteriores. Cerca de 75% das farmácias consideradas apresentam volume de vendas nulo ou muito baixo e existem muito poucas farmácias com grande volume de vendas registadas.
O desvio-padrão assume um valor mais do que duas vezes superior ao valor da média.

			Estatística	Erro Padrão
Check Saúde Pressão Arterial	Média		141,94	7,700
	95% Intervalo de Confiança para Média	Limite inferior	126,84	
		Limite superior	157,04	
	Mediana		3,00	
	Variância		126408,577	
	Desvio Padrão		355,540	
	Mínimo		0	
	Máximo		3758	
	Intervalo		3758	
	Intervalo interquartil		92	
	Assimetria		4,318	,053
	Curtose		23,797	,106





Através da caixa de bigodes verificamos que 50% das farmácias registam entre 0 e 92 unidades vendidas. As restantes 50% são praticamente outliers superiores.

5.1.4 Administração de injetáveis

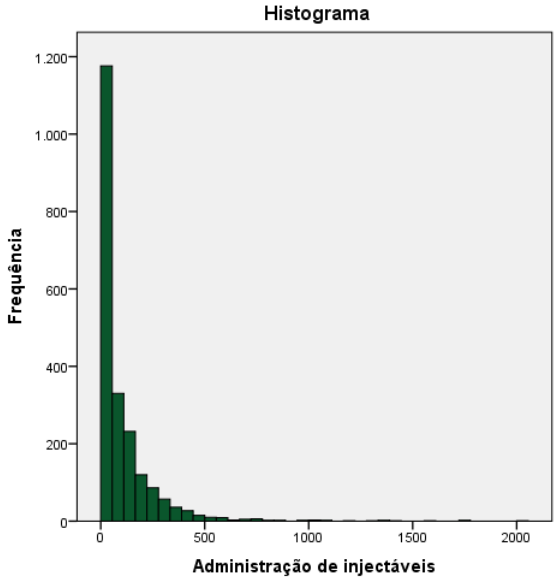
Nos últimos anos as farmácias passaram a poder prestar um vasto conjunto de serviços farmacêuticos, nomeadamente o serviço de administração de vacinas não incluídas no Plano Nacional de Vacinação.

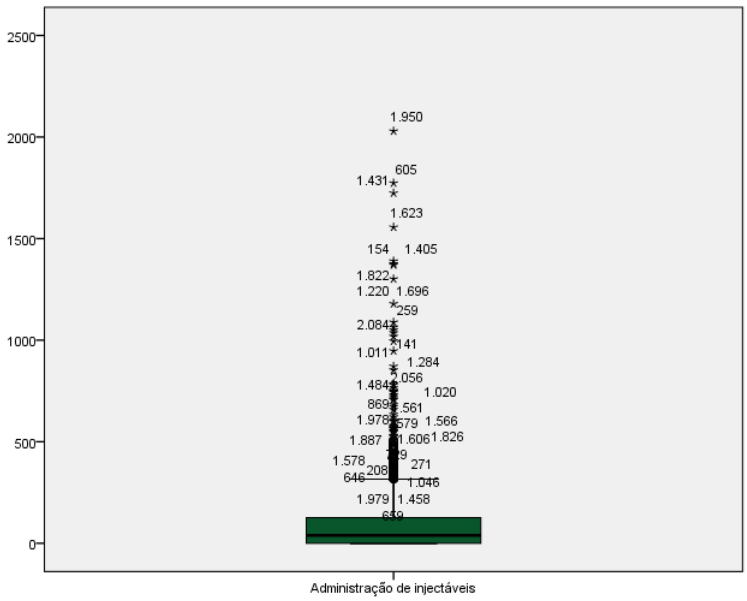
Este serviço é considerado de conveniência para o doente que passa a poder dirigir-se a uma farmácia para usufruir deste serviço, onde a vacina lhe será administrada profissionais legalmente habilitados para o efeito.

Os serviços de administração de injetáveis também se enquadram neste âmbito. Trata-se de administração de medicamentos injetáveis, sejam ou não, vacinas.

Relativamente a este serviço, no que diz respeito à análise descritiva, não há nada de novo a acrescentar.

			Estatística	Erro Padrão
Administração de injectáveis	Média		96,65	3,603
	95% Intervalo de Confiança para Média	Limite inferior	89,58	
		Limite superior	103,71	
	Mediana		39,00	
	Variância		27675,002	
	Desvio Padrão		166,358	
	Mínimo		0	
	Máximo		2029	
	Intervalo		2029	
	Intervalo interquartil		127	
	Assimetria		4,421	,053
	Curtose		30,942	,106





5.2 Construção e validação do modelo

Depois de organizar a base de dados (codificando as categorias de cada uma das características), o primeiro passo foi criar variáveis dummy para identificar a categoria das características nominais.

Algumas características, apesar de quantitativas, foram agrupadas em classes o que nos impedia de as tratar como numéricas. Exemplos dessas variáveis são, o número de montras, o quadro farmacêutico, o quadro pessoal da farmácia, a àrea de atendimento e as dimensões da farmácia.

A característica "Número de montras" estava dividida em "0","1","2" e "mais de 2" e não nos pareceu que fizesse sentido considerá-la ordinal por se ter agrupado numa categoria as farmácias com mais de duas montras. Assim sendo, resolvemos codificar a categoria "mais de 2" com o valor 3 e assumir que uma farmácia que tivesse mais de 2 montras tinha exatamente 3 montras.

Na verdade não devemos estar muito longe da realidade até porque "dificilmente" se encontrará uma farmácia com mais de 3 montras.

Resolvemos agir de modo semelhante com as variáveis "Quadro farmacêutico" e "Quadro pessoal da farmácia".

A situação da variável "Quadro farmacêutico" é idêntica à "Número de montras". Consideramos que a categoria mais de 4 farmacêuticos era exactamente 5. No que diz respeito à variável "Quadro pessoal da farmácia", que estava dividida em classes, considerámos o ponto médio de cada classe. As alterações foram as seguintes:

Variáveis	Antes	Depois
Número	0	0
de	1	1
montras	2	2
	>2	3
Quadro	1	1
farmacêutico	2	2
da	3	3
Farmácia	4	4
	>4	5
Quadro de Pessoal	1-3	2
da	4-6	5
Farmácia	6-10	8
(incluindo o Director-Técnico)	> 10	11

Desta forma, passámos a trabalhar com 11 variáveis categóricas e 3 quantitativas.
Tentar incluir vários níveis de uma variável quantitativa como um fator pode tornar o modelo mais pesado e a interpretação pode não ser tão clara, sendo este o principal motivo para este "ajustamento".

5.2.1 Checksaúde Colesterol Total

Começámos por usar os métodos "Stepwise", "Forward"e "Backward"para uma primeira seleção das variáveis explicativas a considerar no modelo.
Os métodos "Stepwise"e "Forward" selecionaram exatamente as mesmas 4 variáveis (é bastante comum estes dois métodos selecionarem as mesmas variáveis) e o método "Backward" selecionou 6 variáveis (as 4 selecionadas pelos outros métodos mais 2). O acréscimo no R^2 era de $0,01$ e portanto, tendo em conta o princípio da parcimónia, não fazia sentido considerar as 6 características selecionadas pelo método "Backward" se podiamos ter um modelo mais simples (com 4 variáveis) sem alteração significativa no coeficiente de determinação.
Até porque, ainda temos de considerar eventuais interações entre as 4 características.
O estudo das interações foi feito usando a estatística de teste F que enunciámos no capítulo "Modelo completo e modelo reduzido".
Considerámos como modelo completo o modelo com todas as variáveis explicativas selecionadas incluindo as interações duas a duas e o modelo reduzido sem as interações.
Começamos por introduzir todas as interações e usámos uma espécie de metodologia "Backward" ou seja, não eliminámos as interações todas de uma vez e fomos eliminado em cada etapa as interações em que o p-value associada à estatística de teste utilizada era maior. Assim, após algumas iterações, eliminamos a existência de interação entre quaisquer variáveis e portanto concluímos que estaríamos na presença do modelo aditivo.

Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	,325[a]	,105	,103	155,893

a. Preditores: (Constante), Meio_envolvente=Residencial, Instalações=Regulares, Número de Montras, Meio_envolvente=Escritórios / Serviços, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico), Instalações=Antigas

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	6081983,460	6	1013663,910	41,710	,000[b]
	Resíduo	51643242,17	2125	24302,702		
	Total	57725225,63	2131			

a. Variável Dependente: Check Saúde Colesterol Total

b. Preditores: (Constante), Meio_envolvente=Residencial, Instalações=Regulares, Número de Montras, Meio_envolvente=Escritórios / Serviços, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico), Instalações=Antigas

Coeficientes[a]

Modelo		Coeficientes não padronizados	t	Sig.
1	(Constante)	70,846	5,167	,000
	Número de Montras	14,146	2,970	,003
	Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)	20,715	10,112	,000
	Instalações=Antigas	-38,098	-3,960	,000
	Instalações=Regulares	-27,778	-3,567	,000
	Meio_envolvente=Escritórios / Serviços	-33,337	-2,482	,013
	Meio_envolvente=Residencial	-18,707	-2,506	,012

a. Variável Dependente: Check Saúde Colesterol Total

Chegámos ao modelo,

$$Y_k = 70.846 - 38.098\alpha_{1k} - 27.778\alpha_{2k} - 33.337\beta_{1k} - 18.707\beta_{2k} + 14.146\gamma + 20.715\lambda$$

com,

$$\alpha_{1k} = \begin{cases} 1 & \text{se as instalações da farmácia k são antigas} \\ 0 & \text{caso contrário} \end{cases}$$

$$\alpha_{2k} = \begin{cases} 1 & \text{se as instalações da farmácia k são regulares} \\ 0 & \text{caso contrário} \end{cases}$$

$$\beta_{1k} = \begin{cases} 1 & \text{se o meio envolvente da farmácia k é Escritórios/Serviços} \\ 0 & \text{caso contrário} \end{cases}$$

$$\beta_{2k} = \begin{cases} 1 & \text{se o meio envolvente da farmácia k é Residencial} \\ 0 & \text{caso contrário} \end{cases}$$

γ=Número de montras
λ=Quadro de pessoal da farmácia

Apesar do modelo ser significativo, o coeficiente de determinação é muito baixo o que significa que houve um "mau" ajustamento e as previsões decorrentes serão no mínimo "suspeitas".
Vamos fazer a análise dos resíduos para tentar perceber o motivo de um R^2 tão baixo.

Fazendo o teste da normalidade dos resíduos,

H_0 : Os resíduos seguem distribuição normal Vs H_1 : Os resíduos não seguem distribuição normal

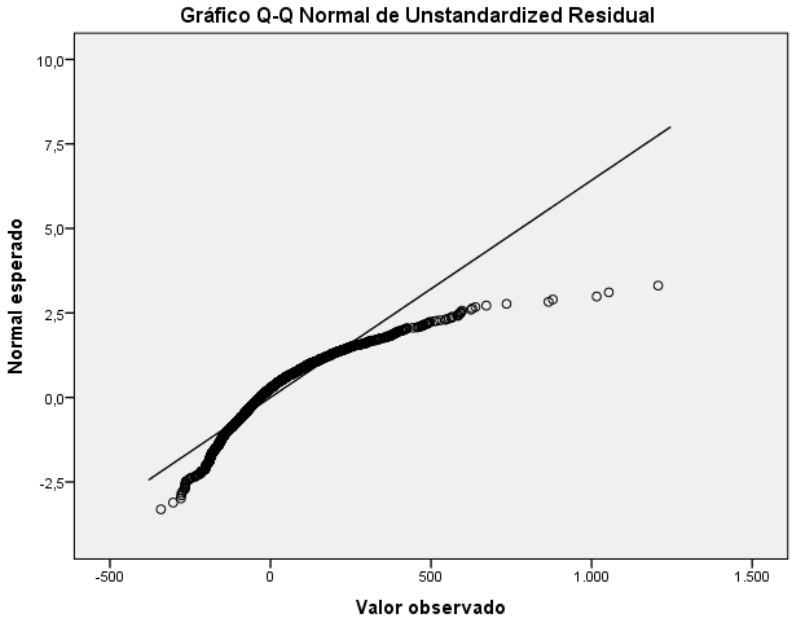
Testes de Normalidade

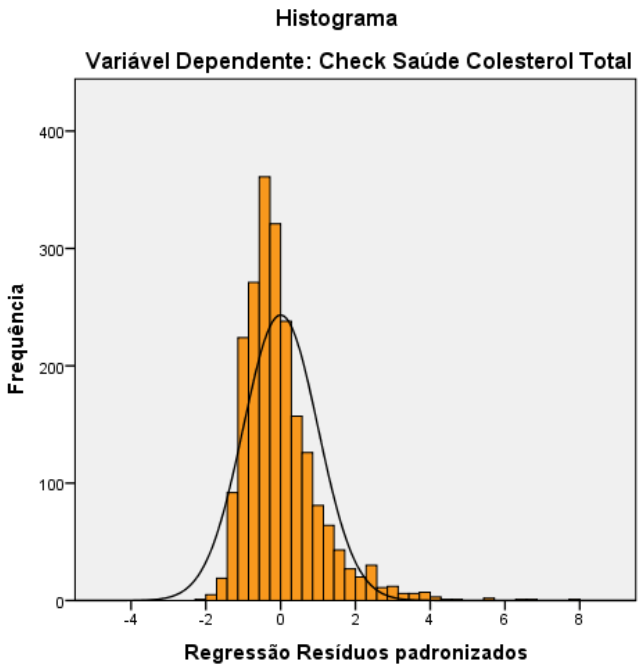
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	df	Sig.	Estatística	df	Sig.
Unstandardized Residual	,117	2132	,000	,879	2132	,000
Standardized Residual	,117	2132	,000	,879	2132	,000

a. Correlação de Significância de Lilliefors

obtém-se um p-value próximo de 0 e portanto rejeitamos a hipótese nula e concluímos que não há evidência estatística para assumir a normalidade dos resíduos e portanto houve violação dos pressupostos. Não é de estranhar este fato se atendermos ao formato exponencial observado no capítulo da análise descritiva.

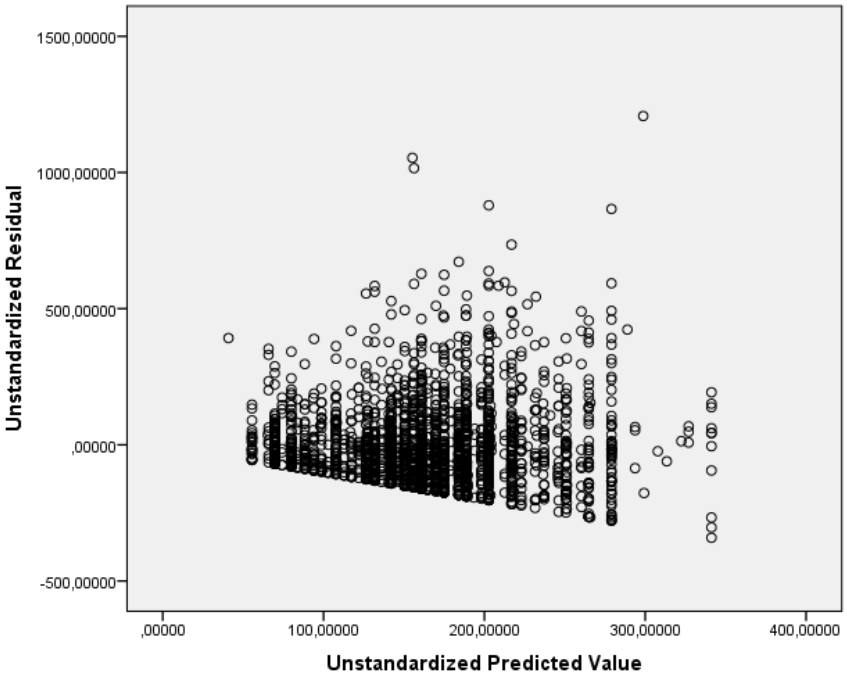
Gráficamente podemos verificar através do qq-plot que os quantis empíricos dos resíduos não se dispõem sobre a bissectriz dos quadrantes ímpares. Existe um mau ajustamento nas caudas que certamente se deve à existência de outliers superiores e ao excesso de observações nulas.





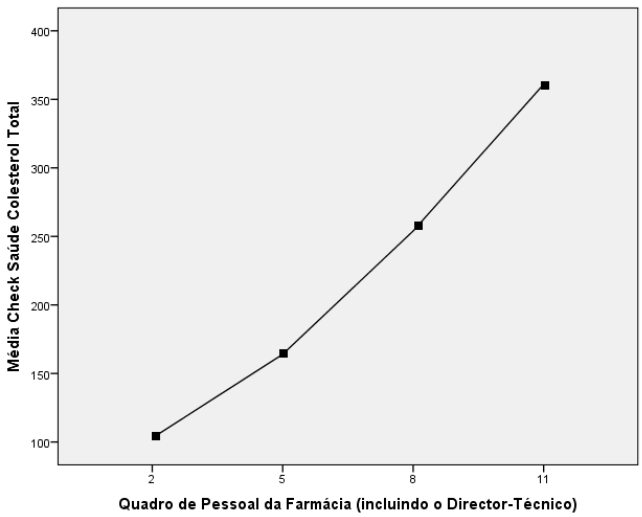
A curva de frequências também não se ajusta à densidade da gaussiana.
Logo, tendo em conta as representações gráficas, não conseguimos validar a hipótese de que os erros seguem distribuição normal.

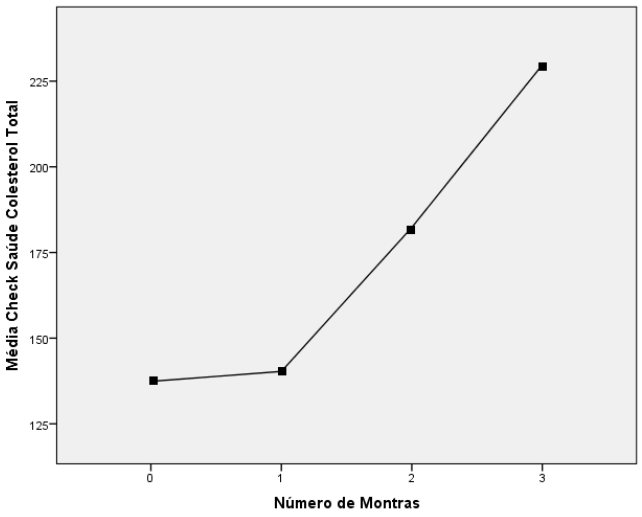
Representando gráficamente Resíduos Vs Valores ajustados



Para que possamos validar a hipótese de que os erros são independentes, de média nula e variância constante os pontos deste gráfico deviam estar aleatoriamente distribuídos em

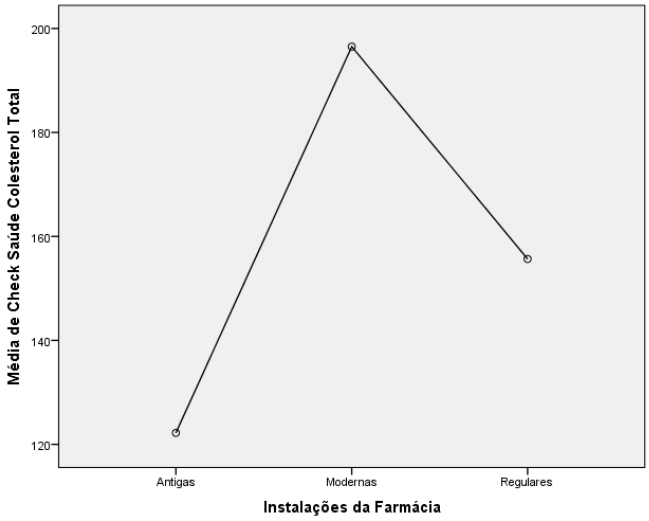
torno da recta que corresponde ao resíduo zero, formando uma mancha de largura mais ou menos uniforme. Se considerarmos apenas os resíduos positivos do gráfico, os pontos parecem dispersar-se de forma aleatória o que não acontece quando olhamos para a parte negativa. Aí, os resíduos parecem estar a dispersar-se segundo um padrão e portanto a condição de independência não é satisfeita. A hipótese de homogeneidade de variância também deve ser posta em causa uma vez que a dispersão dos resíduos parecem apresentar um comportamento tendencialmente crescente.
Os pontos no gráfico parecem sobrepor-se uns aos outros formando uma recta vertical. Isto quer dizer que, farmácias com as mesmas características (consideradas) apresentam valores observados muito diferentes devido à estimativa de σ ser muito alta.

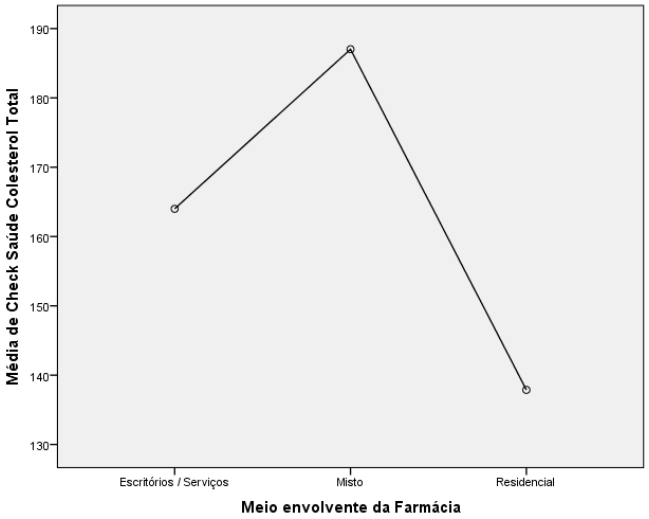




Apesar de não existir uma relação linear forte entre as covariáveis e a variável resposta, através do gráfico acima, ficamos com a ideia de que estas duas características contribuem de certa forma para a variação do volume de vendas que acaba por aumentar com o aumento destes dois fatores. Razão pela qual o coeficiente associado a estas duas covariáveis ser positivo.

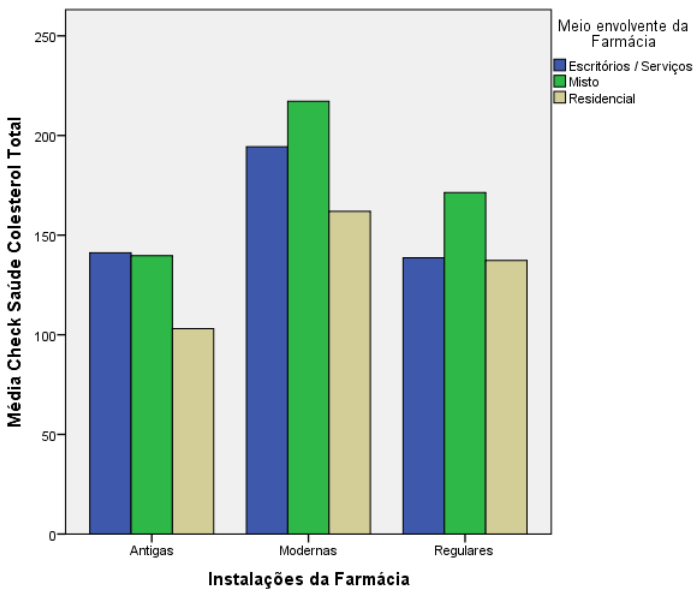
No que diz respeito ás variáveis categóricas consideradas no modelo, elas de fato parecem influenciar a variável resposta e o gráfico abaixo ilustra bem a variação media no volume de vendas.





Farmácias com instalações modernas apresentam um volume de vendas bastante superior às farmácias com instalações antigas.
Em relação à variável "Meio Envolvente" conseguimos facilmente vizualizar que o volume de vendas do serviço Checksaúde Colesterol Total numa farmácia com meio envolvente Misto é bastante superior ao volume de vendas numa farmácia em meio envolvente "Residencial".

No gráfico abaixo podemos verificar que em cada tipo de instalação existe uma certa "uniformidade" do volume médio de vendas relativamente ao fator "Meio Envolvente".
Considerando os dois fatores em simultâneo verifica-se que a frequência máxima (mínima) ocorre nas farmácias com instalações modernas (antigas) e com meio envolvente misto (residencial) que eram as modalidades com frequência marginal máxima (mínima).



Em relação aos restantes serviços analisados neste estudo, procedeu-se de modo completamente análogo chegando-se aos modelos que abaixo apresentamos.
Em qualquer um dos serviços estamos perante um modelo aditivo uma vez que rejeitámos sempre a hipótese de interações significativas entre as variáveis explicativas.

5.2.2 Checksaúde Glicemia

Utilizando o mesmo procedimento utilizado para contruír o modelo para o serviço anterior obtemos:

Coeficientes[a]

Modelo		Coeficientes não padronizados B	t	Sig.
1	(Constante)	12,916	1,627	,104
	Meio_envolvente=Residencial	-11,835	-2,442	,015
	Meio_envolvente=Escritórios / Serviços	-23,548	-2,683	,007
	Número de Montras	14,046	4,725	,000
	Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)	10,884	7,026	,000
	Quadro farmacêutico da Farmácia	7,742	2,715	,007

a. Variável Dependente: Check Saúde Glicemia

MODELO:

$$Y_k = 12.916 - 23.548\beta_{1k} - 11.835\beta_{2k} + 14.046\alpha + 10.884\gamma + 7.742\lambda$$

$$\beta_{1k} = \begin{cases} 1 & \text{se o meio envolvente da farmácia k é Escritórios/Serviços} \\ 0 & \text{caso contrário} \end{cases}$$

$$\beta_{2k} = \begin{cases} 1 & \text{se o meio envolvente da farmácia k é Residencial} \\ 0 & \text{caso contrário} \end{cases}$$

α=Número de montras
γ=Quadro de pessoal da farmácia
λ=Quadro farmacêutico da farmácia

Apesar do modelo construído ser significativo, o coeficiente de determinação assume uma vez mais um valor muito baixo.
Os comentários seriam muito semelhantes à variável anterior pelo que omito.

Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	,308[a]	,095	,093	101,163

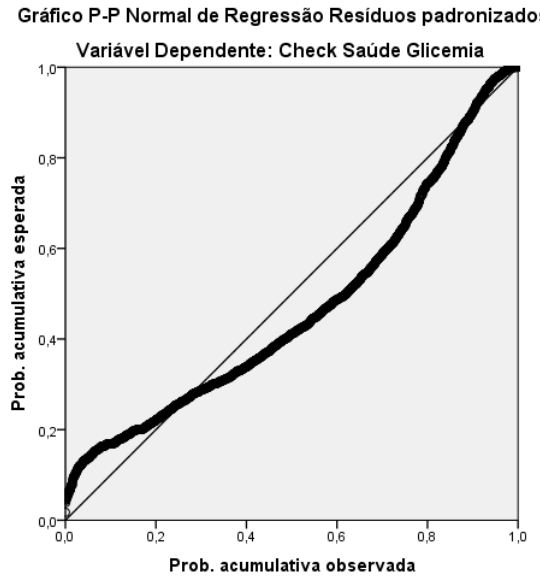
a. Preditores: (Constante), Quadro farmacêutico da Farmácia, Meio_envolvente=Residencial, Meio_envolvente=Escritórios / Serviços, Número de Montras, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)

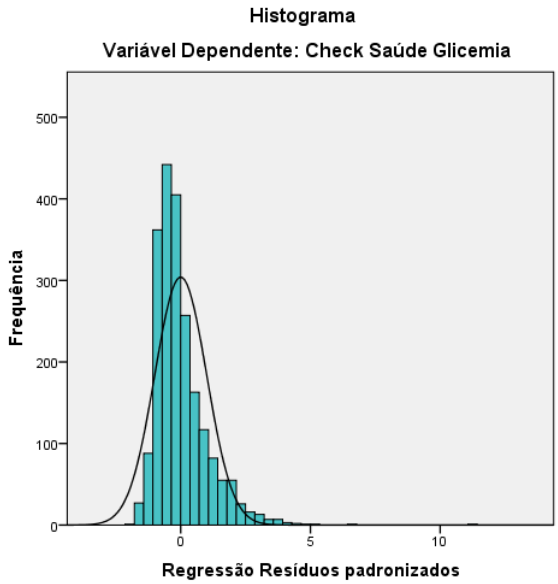
ANOVA[a]

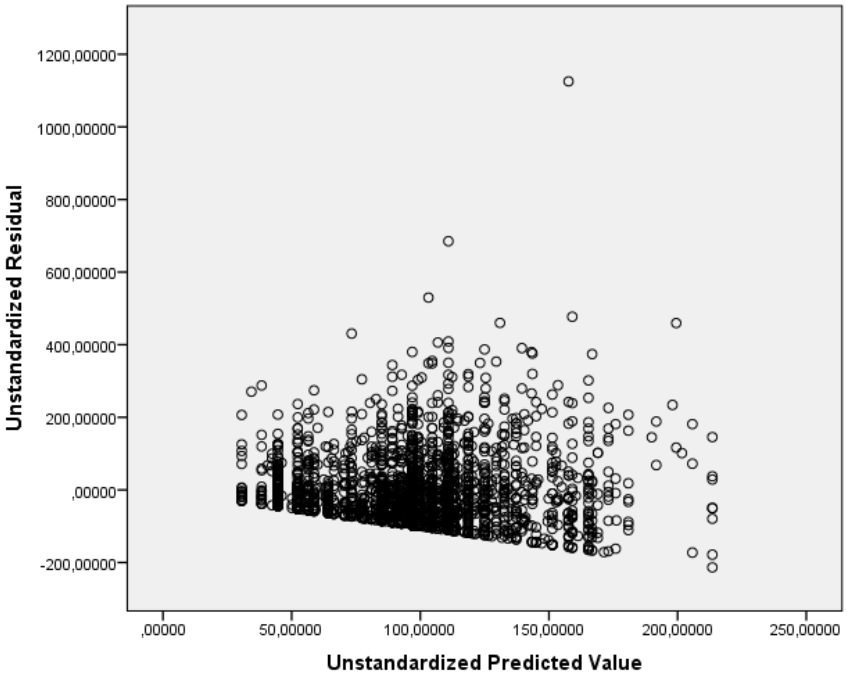
Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	2277977,997	5	455595,599	44,518	,000[b]
	Resíduo	21757178,41	2126	10233,856		
	Total	24035156,40	2131			

a. Variável Dependente: Check Saúde Glicemia

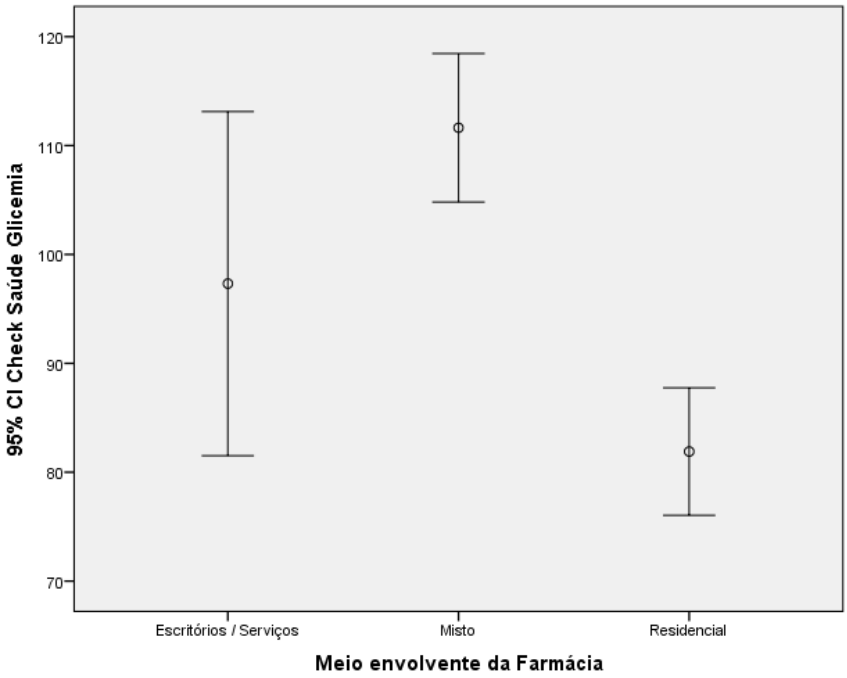
b. Preditores: (Constante), Quadro farmacêutico da Farmácia, Meio_envolvente=Residencial, Meio_envolvente=Escritórios / Serviços, Número de Montras, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)





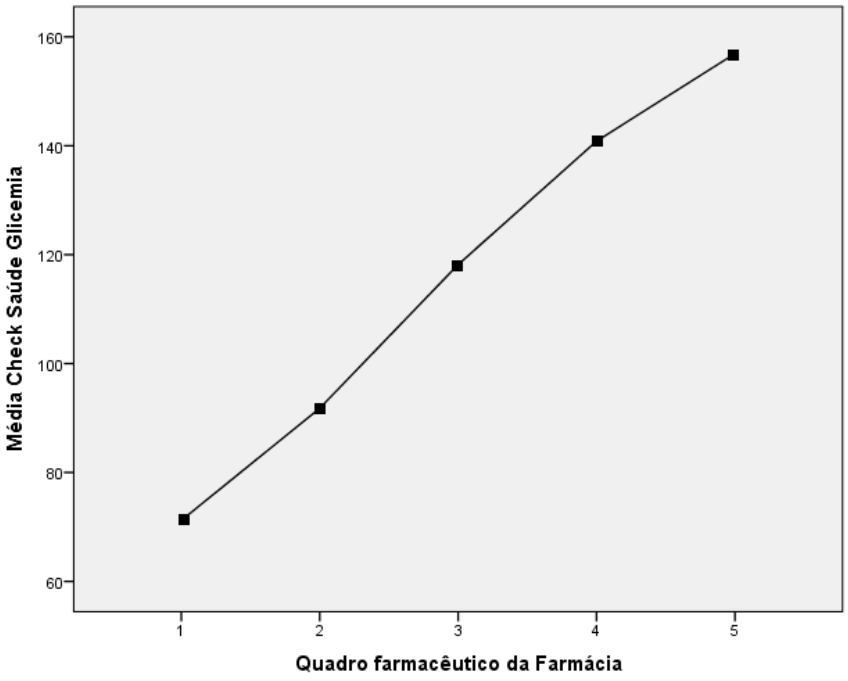


Representando gráficamente o intervalo de confiança a 95% construído para a média do volume de vendas do serviço em estudo para cada uma das categorias da variável "Meio Envolvente"verificamos que



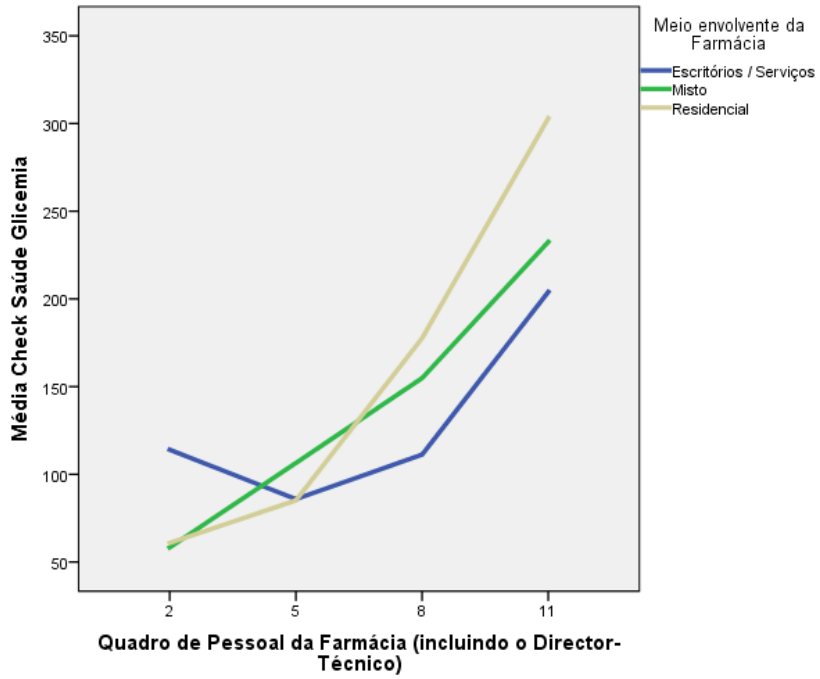
o intervalo de confiança do volume de vendas nas farmácias cujo meio envolvente são Escritórios/Serviços apresenta uma amplitude bastante superior (quase o dobro) dos intervalos de confiança contruídos para as outras duas categorias desta variável.

Mas também conseguimos facilmente vizualizar que o volume de vendas do serviço Check-saúde Glicemia numa farmácia com meio envolvente Misto é bastante superior ao volume de vendas numa farmácia em meio envolvente "Residencial".

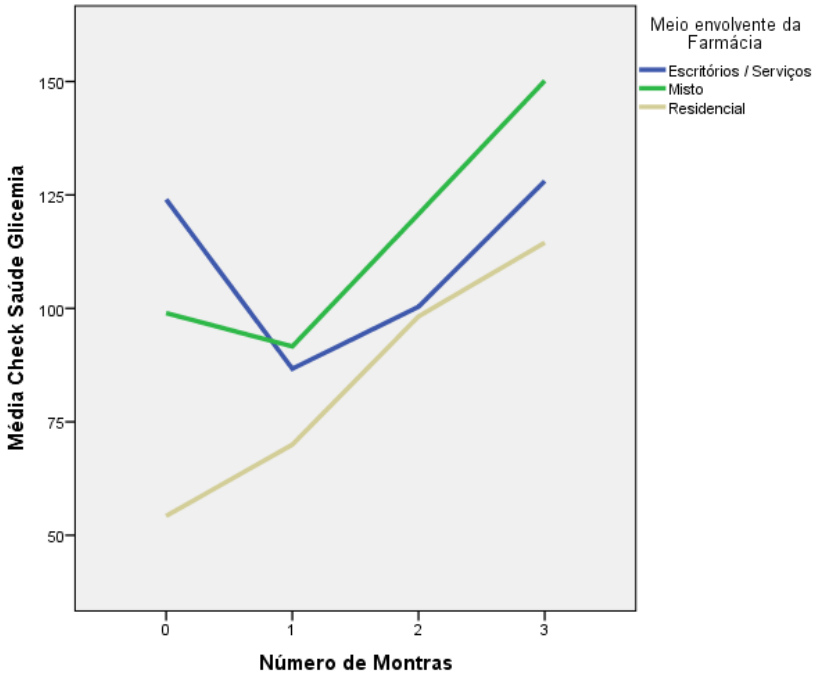


O volume de vendas registadas deste serviço tende a ser maior nas farmácias com maior número de farmacêuticos. Estamos perante uma relação causa efeito-maior volume de vendas, maior quadro de pessoal.

Representando graficamente a variação do volume médio de vendas do serviço com a variação das covariáveis "Quadro de pessoal" e "Número de montras" tendo em conta o fator da característica nominal "Meio envolvente da farmácia" obtemos:



O aumento do quadro de pessoal da farmácia tem maior impacto no volume médio de vendas deste serviço nas farmácias em meio envolvente residencial.



No que diz respeito ao número de montras, o impacto anterior já é maior em meio misto.

5.2.3 Checksaúde Pressão Arterial

No que diz respeito à construção do modelo para este serviço, após a utilização dos mesmos métodos de seleção utilizados nos serviços anteriores, concluímos que as variáveis explicativas são: Número de montras, Quadro farmacêutico e Meio onde a farmácia se insere, sendo que, apesar desta última característica estar subdividida em quatro categorias distintas (Urbano, Semi-urbano, Rural e Zona balnear) só existiam diferenças significativas entre o meio Urbano e os restantes. Os outros três níveis não apresentavam grande diferença no volume médio de vendas registadas. Assim, resolvemos agrupar as 3 categorias e construír uma variável dummy que assume o valor 1 se o meio onde a farmácia se insere for "Urbano" e 0 se o meio for Semi-urbano, Rural ou Zona balnear.

Modelo:

$$Y_k = -33.440 + 33.910\alpha_k + 28.242\gamma + 50.042\lambda$$

onde, γ=Número de montras, λ=Quadro farmacêutico da farmácia e,

$$\alpha_k = \begin{cases} 1 & \text{se o meio onde a farmácia k se insere é Urbano} \\ 0 & \text{caso contrário} \end{cases}$$

Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	,181[a]	,033	,031	349,900

a. Preditores: (Constante), Meio_Insere=Urbano, Número de Montras, Quadro farmacêutico da Farmácia

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	8846049,257	3	2948683,086	24,085	,000[b]
	Resíduo	260530629,1	2128	122429,807		
	Total	269376678,4	2131			

a. Variável Dependente: Check Saúde Pressão Arterial

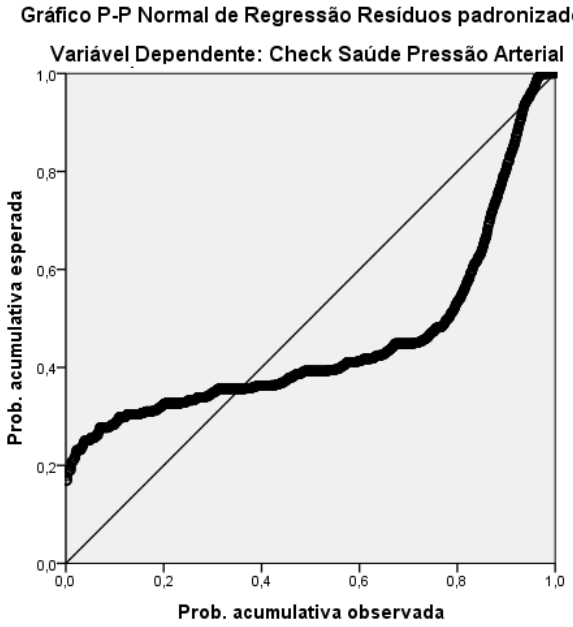
b. Preditores: (Constante), Meio_Insere=Urbano, Número de Montras, Quadro farmacêutico da Farmácia

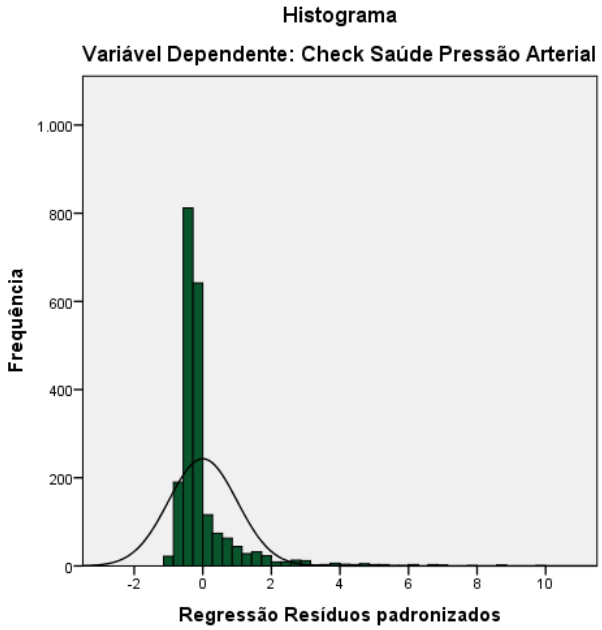
Coeficientes[a]

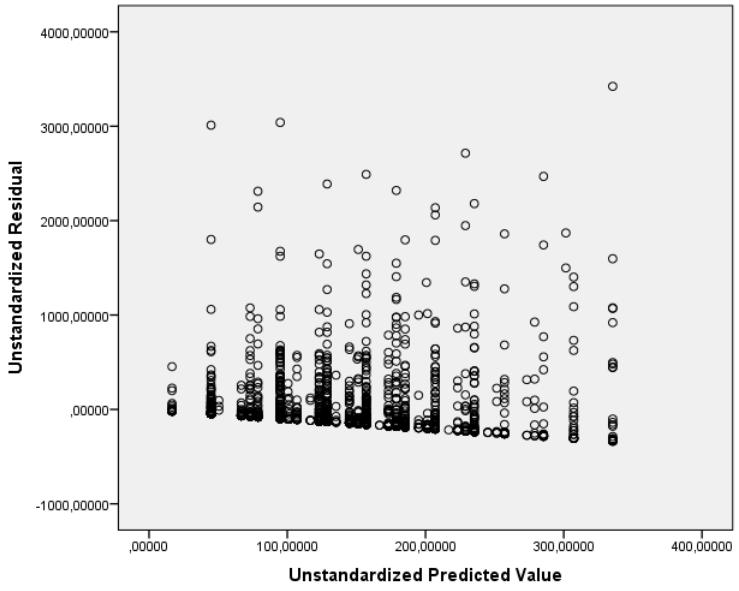
Modelo		Coeficientes não padronizados		Coeficientes padronizados	t	Sig.
		B	Erro Padrão	Beta		
1	(Constante)	-33,440	22,446		-1,490	,136
	Número de Montras	28,242	9,979	,063	2,830	,005
	Quadro farmacêutico da Farmácia	50,042	8,466	,134	5,911	,000
	Meio_Insere=Urbano	33,910	15,898	,048	2,133	,033

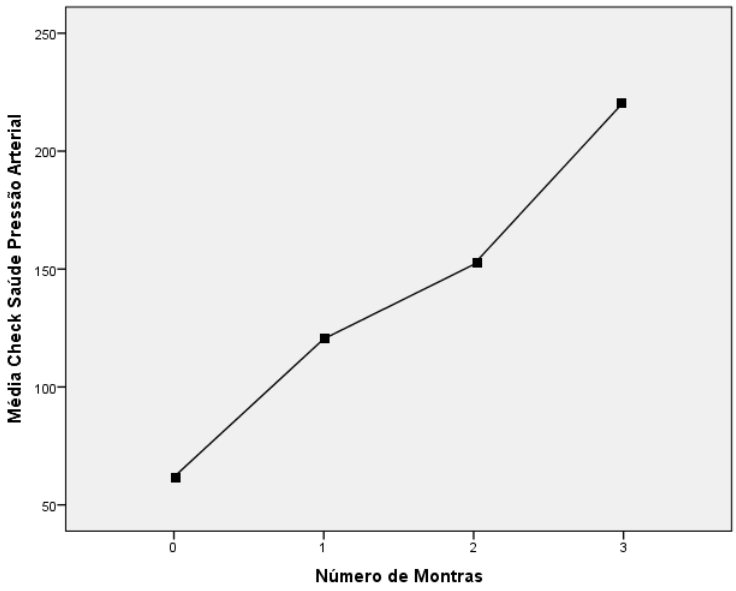
a. Variável Dependente: Check Saúde Pressão Arterial

Apesar do modelo ser signitivativo, este é sem dúvida o modelo com pior ajustamento e portanto o "pior" em termos de previsão.

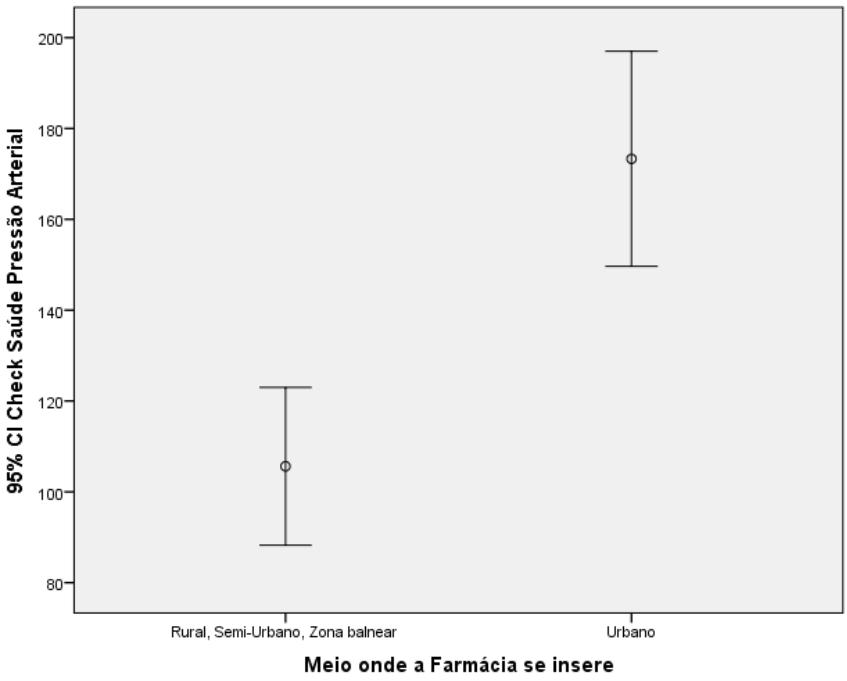




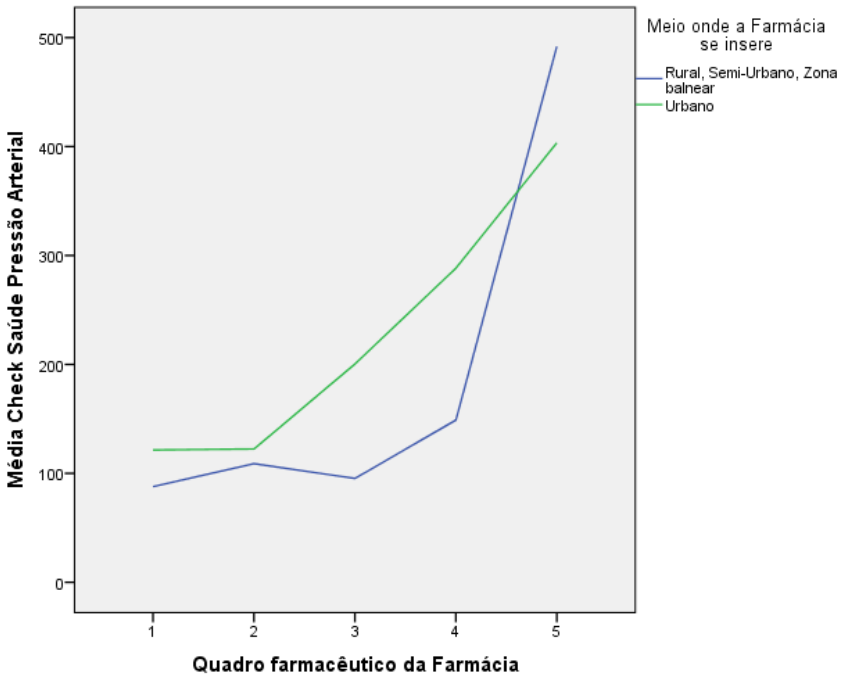




Mais uma vez uma tendência "linear"crescente entre o número de montras e o volume médio de vendas.



Existe uma diferença abismal no volume médio de vendas desta variável entre meio urbano e meio não urbano (rural, semi-urbano, zona balnear).



Apesar do volume médio de vendas registadas ser bastante superior em farmácias no meio urbano, isto não se verifica se o quadro farmacêutico da farmácia for superior a 4 farmacêuticos.

5.2.4 Administração de injetáveis

As variáveis explicativas a considerar para modelar o volume de vendas do serviço Administração de injetáveis são: Quadro pessoal da farmácia, Quadro farmacêutico, Número de montras, Poder de compra dos utentes e Localização da farmácia.
Também aqui achámos que fazia sentido agrupar três categorias da variável Localização numa só e construír uma variável dummy que assume o valor 1 se a farmácia se localiza num Centro Comercial e 0 se a farmácia se localiza numa Rua, Praça grande/Avenida ou Praceta.

Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	,244[a]	,060	,058	161,504

a. Preditores: (Constante), Localização=Centro Comercial, Poder de compra dos utentes, Número de Montras, Quadro farmacêutico da Farmácia, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	3521597,270	5	704319,454	27,002	,000[b]
	Resíduo	55453832,36	2126	26083,646		
	Total	58975429,63	2131			

a. Variável Dependente: Administração de injectáveis

b. Preditores: (Constante), Localização=Centro Comercial, Poder de compra dos utentes, Número de Montras, Quadro farmacêutico da Farmácia, Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados	t	Sig.
		B	Erro Padrão	Beta		
1	(Constante)	-30,388	13,414		-2,265	,024
	Quadro de Pessoal da Farmácia (incluindo o Director-Técnico)	6,117	2,430	,069	2,517	,012
	Quadro farmacêutico da Farmácia	22,965	4,602	,132	4,990	,000
	Número de Montras	12,695	4,770	,060	2,662	,008
	Poder de compra dos utentes	13,317	6,678	,046	1,994	,046
	Localização=Centro Comercial	92,786	34,946	,056	2,655	,008

a. Variável Dependente: Administração de injectáveis

Modelo:

$$Y_k = -30.3880 + 92.785\alpha_k + 6.117\beta + 22.965\gamma + 12.695\lambda + 13.317\theta$$

com,
β=Quadro de pessoal da farmácia,
γ=Quadro farmacêutico da farmácia,
λ=Número de montras,
θ=Poder de compra dos utentes e,

$$\alpha_k = \begin{cases} 1 & \text{se a farmácia se localiza num Centro Comercial} \\ 0 & \text{caso contrário} \end{cases}$$

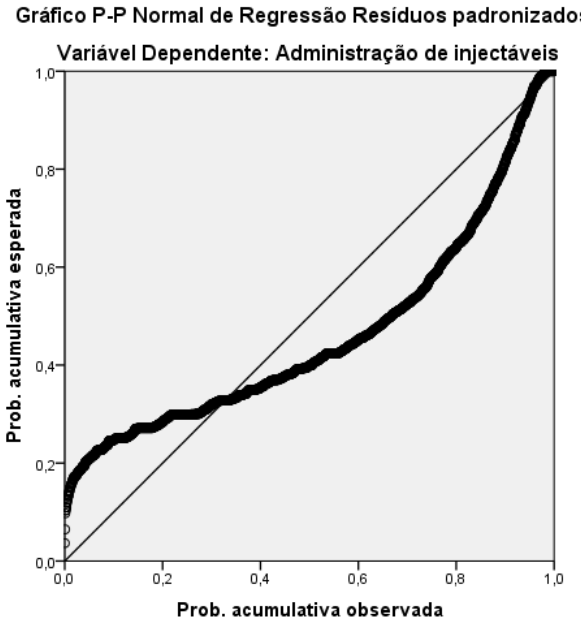
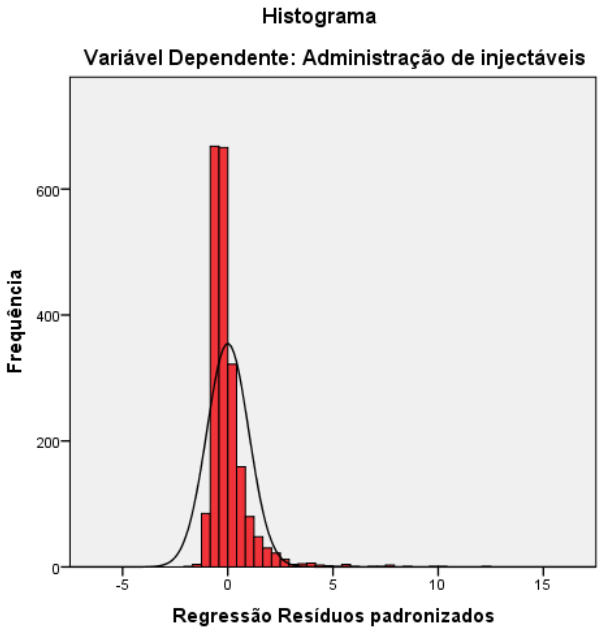
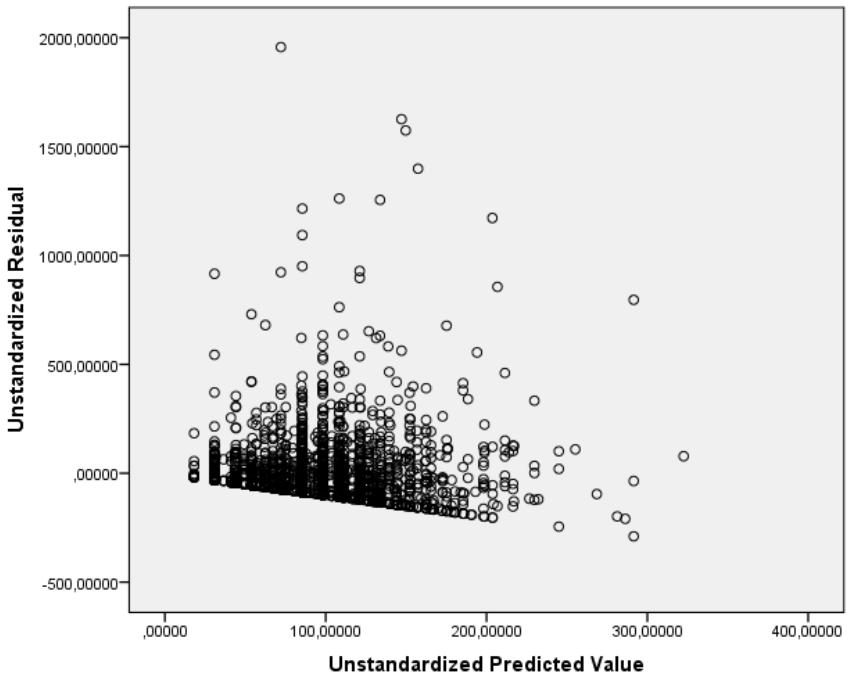
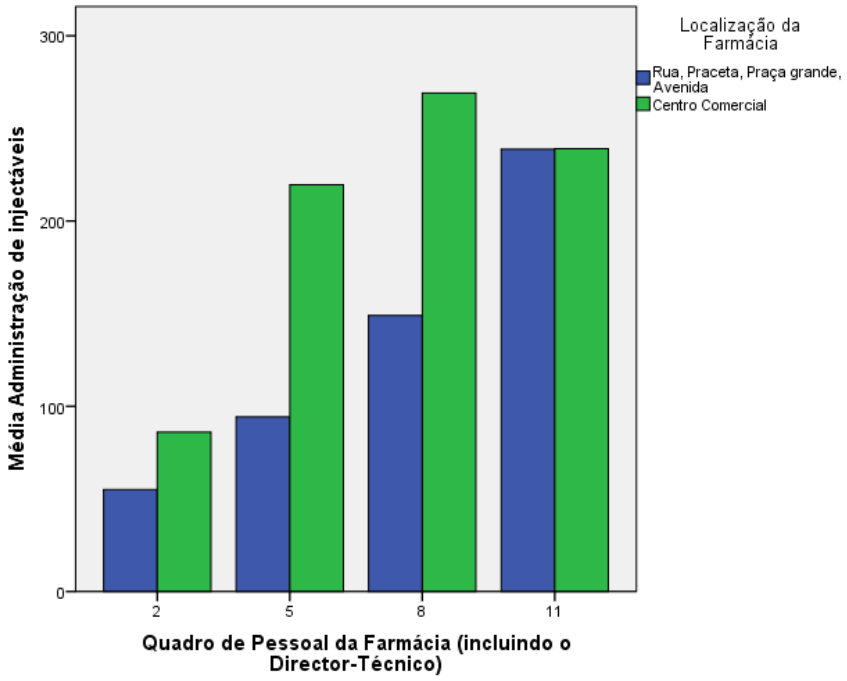
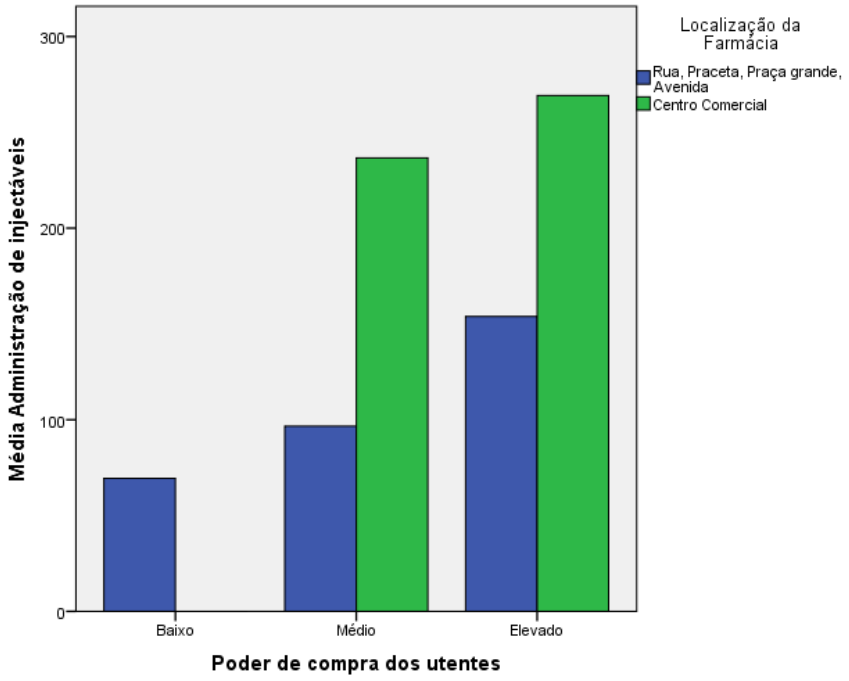


Gráfico P-P Normal de Regressão Resíduos padronizados
Variável Dependente: Administração de injectáveis
Prob. acumulativa esperada
Prob. acumulativa observada
0,0
0,2
0,4
0,6
0,8
1,0


Histograma
Variável Dependente: Administração de injectáveis
Frequência
Regressão Resíduos padronizados
0
200
400
600
-5
0
5
10
15


Unstandardized Residual
Unstandardized Predicted Value
2000,00000
1500,00000
1000,00000
500,00000
,00000
-500,00000
,00000
100,00000
200,00000
300,00000
400,00000





O volume médio de vendas deste serviço é muito superior nas farmácias localizadas em centros comerciais relativamente a outras localizações.

E, manifesta uma tendência crescente significativa com o aumento do poder de compra dos utentes.

De estranhar o facto de não existirem observações desta variável em farmácias localizadas em centros comerciais para utentes com poder de compra baixo.

Capítulo 6

Considerações finais e problemas em aberto

A variável resposta representa o número de vendas registadas de determinado serviço em um ano sendo por isso uma variável discreta mas, o facto de ter uma amplitude de variação tão grande e a dimensão da amostra ser muito elevada faz com que se torne bastante razoável tratá-la como contínua.
No entanto, os pressupostos da normalidade e homocedasticidade dos resíduos não se verificaram; o coeficiente de determinação apresenta um valor muito baixo e, como já referimos, o modelo construído, para cada um dos quatro serviços farmacêuticos considerados, não é o melhor em termos de previsão. Podemos ter maus ajustamentos com modelos estatísticamente significativos. A rejeição da hipótese de nulidade (simultânea ou individual) dos parâmetros das variáveis envolvidas é natural face à dimensão da nossa amostra.
Tendo em conta a distribuição no volume de vendas, em qualquer um dos serviços, apercebemo-nos de uma grande percentagem de valores observados iguais a zero.
O excesso de zeros sugere um modelo misto para a variável resposta. Assim sendo, deveríamos talvez optar por um processo de modelação em duas fases.
Numa primeira fase poderiamos modelar a ocorrência de zeros, recorrendo à regressão logística de forma a separar as farmácias cujo volume de vendas é nulo das farmácias que efetivamente registaram a venda do serviço; depois, procedendo de modo semelhante ao que foi feito, é possível construír um modelo para previsão do volume de vendas nas farmácias que registam venda do serviço.
A segunda fase da modelação terá em conta apenas os valores positivos da variável. Dado que foram registadas vendas de determinado serviço, podíamos tentar modelar o número de ocorrências não nulas.

Seria de esperar que, não considerando as observações nulas, se verificasse um aumento no valor de R^2 mas, experimentámos construír o modelo para um dos serviços (sem os zeros) e na verdade não houve uma melhoria significativa no coeficiente de determinação. Existe uma grande assimetria na distribuição do volume de vendas, o que nos traz a ideia de que poderíamos tentar a distribuição Gamma que é muito usada para modelar distribuições assimétricas que assumam valores positivos ou optar por modelos lineares generalizados no contexto da familia exponencial mas com distribuições assimétricas como por exemplo a distribuição de Weibull ou a Lognormal.

No entanto, a RLM, basesada na normalidade dos termos de erro, é particularmnete robusta no que respeita à distribuição dos mesmos. Para além disso, a amostra disponível era de dimensão razoavelmente grande o que, de acordo com a teoria dos modelos lineares, garante uma boa aproximação à distribuição normal para os EMQ dos coeficientes bem como às estatísticas dos testes T e F.

De notar que os dados fornecidos tinham sido agrupados em classes e o facto de nos terem chegado de forma resumida influencia sem dúvida a análise e portanto, a sugestão que deixamos e que realmente acreditamos trazer uma melhoria na qualidade do ajustamento, seria o não agrupamento das variáveis quantitativas em classes (por exemplo, a área de atendimento, as dimensões, etc).

Outra sugestão seria a operacionalização de certas variáveis com outra escala de medida. Por exemplo, em vez de considerar a característica instalações da farmácia dividida em categorias como antigas, modernas ou regulares poderíamos considerar o número de anos em que a farmácia não sofre qualquer tipo de remodelação.

Uma quantificação mais exata dos preditores leva inevitavelmente a uma melhoria do modelo.

Qualquer modelo que no futuro venha a ser construído, resultado até de possíveis melhorias que sejam implementadas a este, continuarão a ser caricaturas da realidade. Podem eventualmente ser caricaturas um pouco melhores.

Bibliografia

Alpuim, T.,*Modelos lineares*-Notas de apoio à disciplina, 2013.

Maroco, J., Bispo, R., *Estatística Aplicada às Ciências Sociais e Humanas*, Climepsi Editores, 2005.

Cadima, J.,*Modelação Estatística* I, Instituto Superior de Agronomia, 2008.

Pestana, D.D., Velosa, S.F., *Introdução à Probabilidade e à Estatística*, Fundação Calouste Gulbenkian, 2002.

Pestana, M. H. e Gageiro, J.N., *Análise de Dados para Ciências Sociais: A Complementaridade do SPSS*, 4ª Edição, Edições Sílabo, 2005.

Pereira, A., *SPSS - Guia Prático de Utilização Análise de Dados para Ciências Sociais e Psicologia*, 7ª Edição, Edições Sílabo, 2006.

Ferreira, M.C.C.S. (2013): "Modelos de Regressão: uma aplicação em Medicina Dentária" - Tese de Mestrado, Universidade Aberta.

Rodrigues, S.C.A. (2012): "Modelo de Regressão Linear e suas Aplicações" - Tese de Mestrado, Universidade da Beira Interior.