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

Previsão de Consumos de Energia Elétrica em Portugal

Cátia Joana Ribeiro Lourenço

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Prof.ª Doutora Teresa Alpuim

2016

Dedico este trabalho à minha Mãe:

Pela pessoa maravilhosa e extraordinária que ela é e pelos valores que me transmitiu ao longo da sua vida

Agradecimentos

Um especial agradecimento à minha orientadora, Professora Doutora Teresa Alpuim, pela sua ajuda na elaboração deste projeto, pela sua disponibilidade e pelos seus ensinamentos ao longo do Mestrado em Matemática Aplicada à Economia e Gestão.

O tema deste projeto surgiu num estágio profissional na EDP Comercial, a quem desde já agradeço à Direção de Gestão de Energia e Preços não só pela minha formação no setor elétrico, como pelo apoio que obtive durante a elaboração do trabalho de projeto.

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 agradecer a três pessoas que nunca desistiram de mim ao longo destes 5 anos.

Aos meus tios, Ana e Luís, pelo apoio que me forneceram no Ensino Superior e por estenderem sempre a mão quando necessito.

Ao meu namorado, pela motivação, pelo apoio e pelo carinho imensurável durante esta fase académica, pela contribuição para fazer sempre mais e cada vez melhor.

Resumo

Atualmente vivemos num planeta onde a energia elétrica é imprescindível. Diariamente, algumas das nossas rotinas passam por acender luzes, ver televisão, andar de metro/comboio, carregar o telemóvel, entre outras. Mas o que acontece quando se carrega num interruptor e a lâmpada acende? Esta simples tarefa só é possível depois de se ter produzido a energia necessária para esta ação.

Neste projeto começa-se por enquadrar a origem da eletricidade no Mundo e a sua evolução até aos dias de hoje. É também definida a cadeia de valor do setor elétrico em Portugal e os vários consumos de energia elétrica considerados a nível nacional.

Só é possível consumir energia elétrica depois desta ter sido produzida. Uma vez que não existe informação de consumos de cada cliente, surge assim a necessidade de prever o volume de energia para que este possa ser comprado em mercado e, posteriormente, produzido.

Este projeto estuda modelos de regressão linear múltiplos que descrevam o consumo de energia elétrica em Portugal através de variáveis históricas de consumos de energia, temperaturas diárias e velocidade média diária do vento. É de salientar que as variáveis referentes a consumos históricos de energia elétrica têm uma especial atenção uma vez que seguem determinados pressupostos, ao contrário das restantes.

Para ser possível compreender este estudo, alguns métodos estatísticos são descritos e explicados, bem como os vários pressupostos assumidos. Após a definição das variáveis a serem utilizadas em estudo, é feita uma análise preliminar, bem com o estudo de multicolinearidade de cada uma delas. São também utilizados diferentes métodos de seleção de variáveis na criação de modelos de regressão linear múltipla para garantir que sejam obtidos os modelos com melhores resultados.

De modo a ser mais percetível o modelo que fornece melhores resultados, é realizado um teste a cada um dos modelos com dados que não fazem parte do histórico considerado.

Palavras-chave: Modelos de Regressão Linear Múltipla, Variáveis Condicionadas, Multicolinearidade, Consumo de Energia Elétrica

Abstract

Currently we live in a planet where the eletric power is essential. Daily, some of our routines pass for lighting light, to see television, floor of underground/train, to load cell phone, among others. But what happens when it loads an interruptor and the light bulb lights? This simple task is possible after if having produced the necessary energy for this action.

In this project, it is started with the origin and the evolution of the electricity in the world until today. It is also deined the chain of value of the electric sector in Portugal and the some consumptions of electric power considered the national level.

It is only possible consume eletric power after this having been produced. A time that does not exist information of consumptions of each customer, arises the necessity to predict the volume of energy so that this can be bought in market and, later, produced.

This project studies multiple models of regression linear that describe the consumption of the electric power in Portugal trough historical variables of the consumptions of power, daily temperatures and daly average speed of the wind. It is important referring that variables dependentes of consumption has a special treatment, because of their suppositions, in contrast of the remains.

To be possible to understand this study, some statistical methods described and are explained, as well as the several estimated assumed. After the definition of the variable to be used in study, it is made a preliminar analysis, well with the study of multicollinearity of each one of them. Also different methods of election of variable in the creation of models of multiple linear regression are used to guarantee that the resulted models with better are gotten.

In order to be perceptible the model that it supplies better resulted, it is carried through a test to each one of models with dates that are not part of historical.

Keywords: Models of Multiple Linear Regression, Conditional Variables, Multicollinearity, Consumption of eletric Power

Índice

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Siglas e Acrónimos

°C – Graus Celsius
AT – Alta Tensão
BT – Baixa Tensão
BTE – Baixa Tensão Especial
BTN – Baixa Tensão Normal
Cons. - Consumo
CPE – Código Ponto Entrega
EMQ – Estimador de Mínimos Quadrados
EQM – Erro Quadrático Médio
ERSE – Entidade Reguladora dos Serviços Energéticos
km/h – Quilometro por hora
kV - Quilovolt
kVA – Quilovolt-ampere
kW – Quilowatt
kWh – Quilowatt hora
MAT – Muito Alta Tensão
Med. - Média
MT – Média Tensão
REN – Rede Energética Nacional
RESP – Rede Elétrica de Serviço Público
RLM – Regressão Linear Múltipla
RNT – Rede Nacional de Transporte
SEN - Sistema Elétrico Nacional
SQ – Soma dos Quadrados
Temp. - Temperatura
Vel. - Velocidade

Lista de Definições

Alta Tensão – Tensão entre fases cujo valor eficaz é superior a 45kV e igual ou inferior a 110kV
Baixa Tensão – Tensão entre fases cujo valor eficaz é igual ou inferior a 1kV
Baixa Tensão Especial – Fornecimentos ou entregas em BT com potência contratada superior a 41,4kW
Baixa Tensão Normal – Fornecimentos ou entregas em BT com potência contratada igual ou inferior a 41,4kW
Contagem – Medição de energia elétrica num período de tempo determinado
Média Tensão – Tensão entre fases cujo valor eficaz é superior a 1kV e igual ou inferior a 45kV
Muito Alta Tensão – Tensão entre fases cujo valor eficaz é superior a 110kV
Telecontagem - Contagem com leitura remota

Capítulo 1

Introdução

O simples facto de chegar a casa depois de um dia de trabalho e poder ligar as luzes dos candeeiros só é possível depois da energia elétrica ter sido produzida. Com isto surgiu a necessidade de prever o volume de energia para que este possa ser, posteriormente, produzido.

O objetivo deste projeto é a obtenção de boas previsões para os consumos diários de energia elétrica, recorrendo a modelos estatísticos desenvolvidos a partir da regressão linear múltipla, com base em variáveis históricas. Estas variáveis são os consumos históricos de energia elétrica, as temperaturas máximas, mínimas e médias diárias e a velocidade média diária do vento, em Portugal Continental, desde 2011 até 2015.

A construção dos modelos de regressão obtidos tem em conta dois pressupostos: variáveis internas ao consumo de energia elétrica e variáveis externas ao consumo de eletricidade. O primeiro modelo de regressão criado depende apenas de variáveis internas ao consumo, isto é, variáveis que descrevem o consumo desde há 3 até há 7 dias atrás, o consumo anual, o consumo mensal e o consumo semanal. A criação do segundo modelo utiliza, para além das variáveis anteriormente definidas, variáveis externas ao consumo, tais como a temperatura máxima, mínima e média diária e a velocidade média do vento em Portugal.

A análise final deste projeto procura perceber, tendo em consideração o bom ajustamento e a qualidade das previsões, qual é o melhor modelo que descreve os consumos futuros e se depende ou não de variáveis externas ao consumo de energia elétrica.

Enquadramento

O início da eletricidade surgiu no século VI a.C., na Grécia Antiga, com o filósofo *Thales de Mileto*. Este descobriu uma resina vegetal fóssil petrificada, denominada de âmbar. Após esfregá-la em pele e lã de animais observou que pedaços de palhas e fragmentos de madeira começaram a ser atraídos pelo âmbar. Este poder atrativo deu início ao estudo de uma nova ciência.

Em 1600, o médico *William Gilbert* nomeou o evento de atração dos corpos por eletricidade e denominou os objetos que ao serem atritados se eletrizam como objetos elétricos.

Em 1672, *Otto von Guericke* inventou uma máquina geradora de cargas elétricas, através da constante rotação de uma esfera de enxofre causando atrito em terra seca.

Durante o século XVII, a evolução das máquinas elétricas deu origem a um disco rotativo de vidro que era atritado a um isolante adequado. Desta evolução, uma das mais importantes foi o condensador descoberto por *Ewald Georg von Kleist* e por *Petrus van Musschenbroek*. Este condensador consistia numa máquina armazenadora de cargas elétricas. Outra invenção muito importante foi o uso de para-raios realizada pelo físico e político *Benjamin Franklin*.

A descoberta da eletrificação por contato de corpos eletrizados em corpos neutros surgiu mais tarde, em 1730, através do físico *Stephen Gray*. Com esta descoberta surgiu a possibilidade de canalizar a

eletricidade e transportá-la de um corpo para o outro, iniciando-se assim os conceitos de condutores e isolantes elétricos através de materiais que conduzem a eletricidade com maior ou menor eficácia.

No século XVIII, a partir do modelo primitivo de *Otto von Guericke*, *Charles Du Fay* aprofundou as propriedades elétricas de diversos materiais, comprovando também a existência de dois tipos de força elétrica: uma de atração e outra de repulsão. Com isto inicializou-se o conceito de eletricidade vítrea e resinosa. Entende-se por eletricidade vítrea quando o material de referência corresponde à carga positiva e eletricidade resinosa quando o material contém carga negativa.

A existência de dois tipos de eletricidade foi também comprovada de forma independente pelo cientista *Benjamin Franklin*. Em 1750, *Franklin* afirmou que a eletrificação de dois corpos que causavam atrito era a falta de um dos dois tipos de eletricidade num deles.

Ainda no século XVIII, obteve-se pela primeira vez uma fonte de corrente elétrica estável. Posteriormente, foram realizadas experiências de decomposição de água, onde *Humphry Davy* separou eletronicamente o sódio e o potássio.

A ligação entre o magnetismo e a eletricidade foi descoberta pelo físico *Christian Örsted* ao visualizar como um fio de corrente elétrica age sobre uma agulha de uma bússola.

Em 1831, *Michael Faraday* descobre que através de um circuito fechado é induzida uma corrente numa bobina através da variação da intensidade na corrente elétrica. Deste modo, conclui-se que a corrente também poderia ser produzida através de magnetismo. Uma bobina próxima de um íman é um exemplo de um gerador de corrente elétrica alternada.

O aperfeiçoamento dos geradores foi um ponto essencial no suprimento da eletricidade aplicada na iluminação. A fim de estimular a invenção de turbinas a vapor pela utilização de energia hidroelétrica, foram criadas máquinas a vapor para movimentar os geradores. Em 1886 foi construída a primeira hidroelétrica junto das Cataratas do Niágara.

A partir de 1885, *Heinrich Hertz* com uma das suas experiências deu início ao estudo das propriedades das ondas eletromagnéticas geradas por uma bobina de indução. Com este estudo demonstrou que as ondas de rádio e de luz são ambas ondas eletromagnéticas. Este estudo vai de encontro com as teorias de *Maxwell* ao confirmar que duas ondas apenas diferem na sua frequência.

A partir do século XIX foi aceite a proposta de que a eletricidade é um conjunto de reações de partículas elementares do átomo (protões, eletrões e neutrões), em que esta se manifesta através da atração ou repulsa entre as cargas. Deste modo, a eletricidade pode ser originada a partir de cargas elétricas, quer em repouso denominada de eletricidade estática, quer em movimento denominada de corrente elétrica.

De modo geral, a eletricidade é conhecida como sendo apenas energia elétrica, onde esta é gerada através de diferenças de potencial elétrico entre dois pontos e permitem estabelecer uma corrente elétrica entre ambos.

A eletricidade pode gerar calor, movimento ou luz e produzir-se de forma natural, através das descargas elétricas dos raios durante as trovoadas.

Nos dias de hoje, a eletricidade é considerada um bem essencial para o Homem, uma vez que é utilizada para pôr a funcionar todo o tipo de máquinas, dispositivos eletrónicos e sistemas de transporte no dia-a-dia.

Capítulo 2

2. Eletricidade em Portugal

2.1 Cadeia de Valor do Setor Elétrico

Nos dias de hoje, o Homem encontra-se, para ligar simplesmente o candeeiro ou até mesmo carregar o telemóvel, dependente de energia elétrica. Para que estas ações ocorram é necessário produzir energia, transportá-la e, posteriormente, distribuí-la até às habitações/instalações do consumidor. Este percurso define a cadeia de valor do Setor Elétrico: produção, transporte, distribuição e comercialização, ilustrado na figura 2.1.

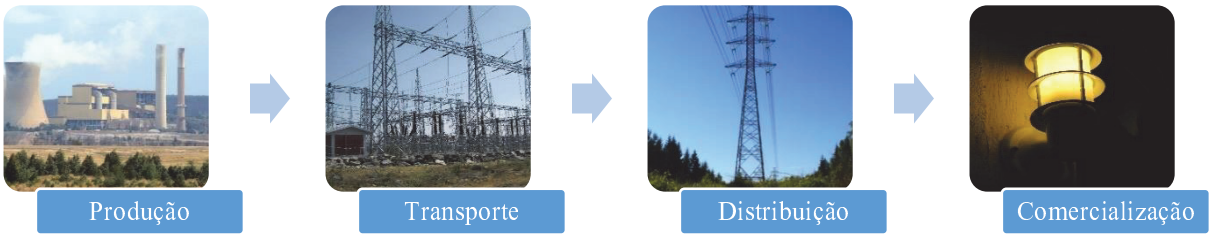

Figura 2.1: Descrição da Cadeia de Valor do setor elétrico

2.1.1 Produção

A energia elétrica pode ser produzida através de diversas fontes renováveis e não renováveis de energia. Define-se fonte de energia renovável como sendo aquelas em que a sua utilização e uso é renovável, ou seja, são fontes de energia inesgotáveis ou que podem ser repostas a curto ou médio prazo, espontaneamente ou por intervenção humana. Por outo lado, fontes não renováveis têm recursos limitados, sendo que esse limite depende dos recursos existentes no planeta.

De entre as fontes primárias poderão ser utilizadas na produção de energia elétrica o carvão, a hídrica, a eólica, a solar, a geotérmica, as marés, as ondas, a biomassa, o petróleo, o gás natural e o urânio.

A energia hídrica é uma fonte de energia renovável e é obtida a partir do curso de água, sendo também aproveitada por um desnível ou queda de água. Por outro lado, a fonte de energia denominada de marés é obtida através do movimento de subida e descida do nível de água do mar, enquanto a fonte de energia através das ondas consiste no movimento ondulatório das massas de água, por efeito do vento.

A biomassa trata-se do aproveitamento energético da floresta e dos seus resíduos, dos resíduos da agropecuária, da indústria alimentar ou dos resíduos do tratamento de efluentes domésticos e industriais. A partir desta fonte de energia pode-se produzir biogás e biodiesel.

A energia eólica e a energia solar provêm, respetivamente, do vento e da luz do sol, enquanto a fonte de energia geotérmica provém do aproveitamento do calor do interior da Terra.

O carvão, petróleo, gás natural e urânio são fontes de energia não renováveis.

O gráfico 2.1[1] representa a contribuição de cada fonte usada na produção de energia elétrica desde 2010 a 2015.

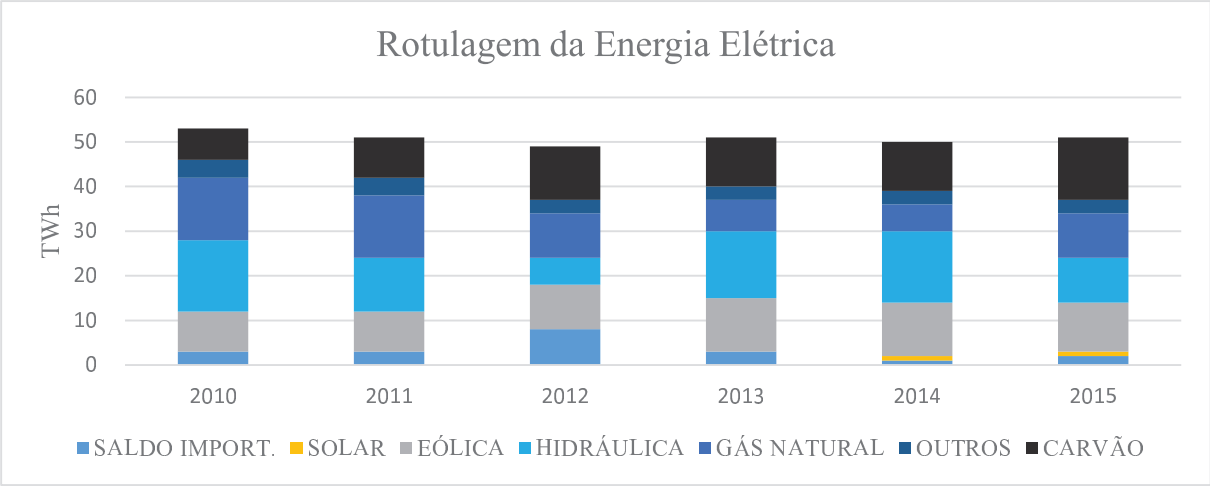

Gráfico 2.1: Evolução das fontes primárias usadas na produção de energia elétrica de 2010-2015

Ao longo dos 6 anos, o uso de carvão na produção de energia elétrica aumentou consideravelmente, enquanto o uso de eólica estagnou. Por outo lado, o uso de energia solar na produção teve impacto apenas a partir de 2014. As restantes fontes de energia não seguem um padrão ao longo dos 6 anos.

Durante o ano 2015, as fontes renováveis representaram 47% da produção total de energia, equiparada com as energias não renováveis, 48%, como se poderá confirmar pelo gráfico 2.2[2].

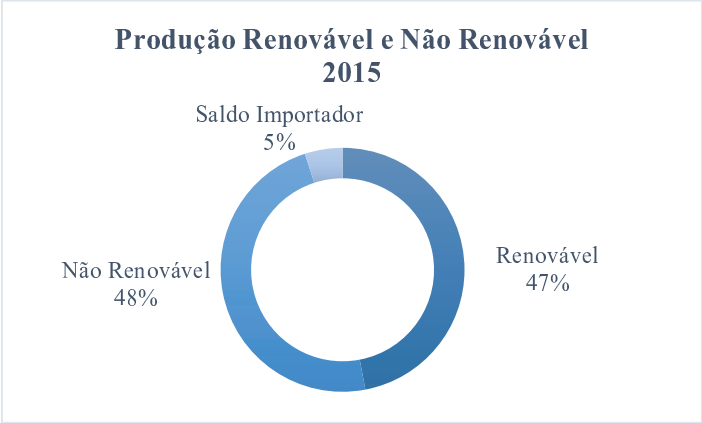

Gráfico 2.2: Distribuição das fontes renováveis e não renováveis usadas na produção de energia elétrica em 2015

2.1.2 Transporte

O transporte de eletricidade dos centros produtores até aos centros de consumo é realizado em linhas de Muito Alta Tensão (MAT).

O transporte de energia elétrica é realizado pela Rede Nacional de Transporte (RNT), mediante uma concessão concedida pelo Estado Português, em regime de serviço público. Esta concessão, em Portugal, é da exclusividade da Rede Energética Nacional (REN), onde esta inclui o planeamento, a construção,

[1] Fonte: Relatório e Contas 2015, REN
[2] Fonte: Relatório e Contas 2015, REN

a operação e manutenção da RNT, contendo ainda o planeamento e a gestão técnica do Sistema Elétrico Nacional (SEN), para assegurar o bom funcionamento das infraestruturas, a continuidade de serviço e a segurança do fornecimento de eletricidade.

A energia transportada na RNT desde 2011 até 2015 encontra-se no gráfico 2.3[3], tal como a evolução do consumo no mesmo período de tempo.

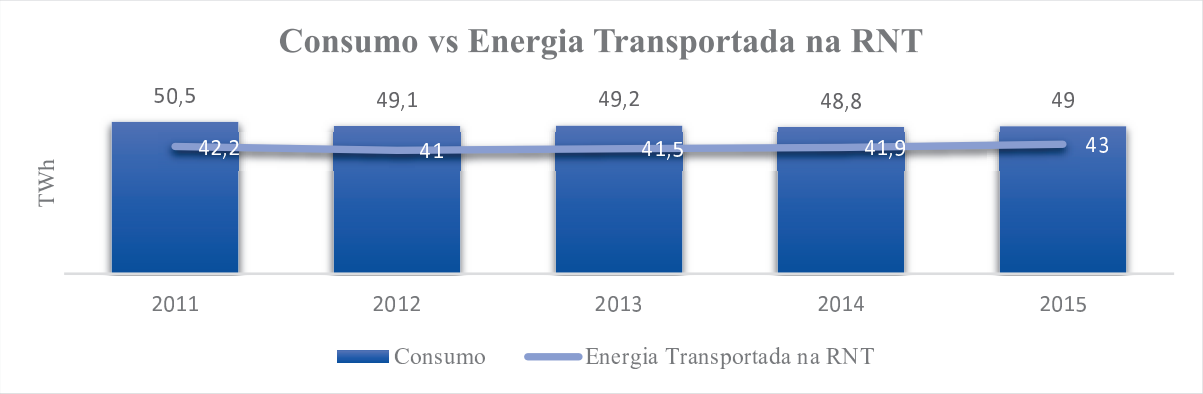

Gráfico 2.3: Evolução do consumo e da Energia Transportada na RNT, em TWh, ao longo de 5 anos

A diferença anual entre o consumo e a energia transportada na RNT deve-se não só às perdas existentes na rede de transporte, como a produção de eletricidade para autoconsumo, isto é, produção de energia elétrica destinada à satisfação de necessidades próprias de abastecimento de energia elétrica do produtor, sem prejuízo do excedente de energia ser injetado na Rede Elétrica de Serviço Público (RESP).

A entrada e a saída de energia elétrica na rede podem ser feitas através de centro produtores, de rede de distribuição e de interligações. As interligações são as ligações por uma ou várias linhas entre duas ou mais redes, sendo possível deste modo a importação e a exportação de energia elétrica. Por fim, as perdas na rede é a diferença entre a energia elétrica que entra na rede e a energia elétrica que sai.

Nos anos 2014 e 2015, a distribuição da energia que entrou e saiu na rede encontra-se descrita na tabela 2.1[4].

Tabela 2.1: Distribuição das entradas e saídas de energia na rede e as perdas nos anos 2014-2015, em TWh

TWH	2014	2015
ENERGIA ENTRADA NA REDE	**41,9**	**43,0**
CENTRO PRODUTORES	32,2	33,0
INTERLIGAÇÕES	7,2	8,1
REDE DE DISTRIBUIÇÃO	2,5	1,9
ENERGIA SAÍDA DA REDE	**41,1**	**42,3**
CENTRO PRODUTORES/CLIENTES DIRETOS	2,9	3,3
INTERLIGAÇÕES	6,4	5,8
REDE DE DISTRIBUIÇÃO	31,8	33,2
PERDAS	**0,8**	**0,7**

2.1.3 Distribuição

[3] Fonte: Relatório e Contas 2011, Relatório e Contas 2013, Relatório e Contas 2015 - REN
[4] Fonte: Relatório e Contas 2015, REN

Para além do transporte em MAT é necessário fazer a distribuição da eletricidade através das redes de distribuição em Alta, Média e Baixa Tensão (AT, MT e BT), onde se encontram os consumidores finais. A atividade de distribuição em Portugal Continental é efetuada maioritariamente pela EDP Distribuição e por algumas cooperativas de distribuição de energia elétrica em BT.

As redes de distribuição de BT são operadas ao abrigo do acordo de concessão firmada mediante concurso público lançado pelos municípios.

O gráfico 2.4[5] demonstra a evolução do volume de energia, em GWh, perdido na rede ao longo de 5 anos, desde 2011-2015 e a evolução do comprimento das linhas, em quilómetros.

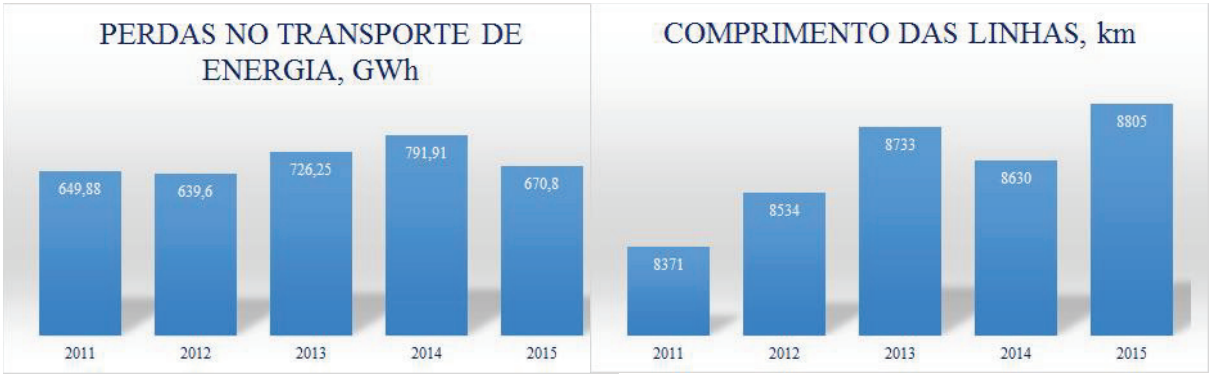

Gráfico 2.4: Evolução das perdas no transporte de energia e no comprimento das linhas de 2011-2015

2.1.4 Comercialização

A comercialização é compreendida pela compra e venda de eletricidade, onde os comercializadores têm o direito de aceder às redes de transporte e de distribuição mediante o pagamento de tarifas de acesso. Estas tarifas são estabelecidas pela Entidade Reguladora dos Serviços Energéticos (ERSE). O serviço público dos comercializadores tem não só direitos como obrigações. Estes têm a obrigação de disponibilizar informação de forma simples e compreensível aos seus clientes e de garantir a qualidade e continuidade no abastecimento de eletricidade.

É também da responsabilidade dos comercializadores a faturação e o atendimento ao cliente. Em Portugal continental existem dois tipos de comercializadores: os comercializadores livres e os comercializadores de último recurso, onde a comercialização é feita, respetivamente, em Mercado Livre e em Mercado Regulado.

O Mercado Livre é definido como sendo o mercado em que os comercializadores de eletricidade concorrem livremente entre si em preços e condições comerciais, respeitando assim as regras de concorrência e o Regulamento das Relações Comerciais. Neste tipo de mercado, os consumidores têm o direito de escolher livremente o seu comercializador de energia. Por outro lado, no mercado regulado, os preços de venda de energia, para os consumidores finais, são fixados anualmente pela ERSE.

O consumo de energia elétrica no Mercado Livre tem vindo a aumentar consideravelmente a partir de 2010 e, em contrapartida, o consumo no Mercado Regulado a diminuir. Isto deve-se ao elevado número

[5] Fonte: Relatório e Contas 2015, REN

de clientes transitarem do Mercado Regulado para o Mercado Livre. O gráfico 2.5[6] representa a evolução do consumo anual, desde 2010 a 2015, em Portugal no Mercado Livre e no Mercado Regulado.

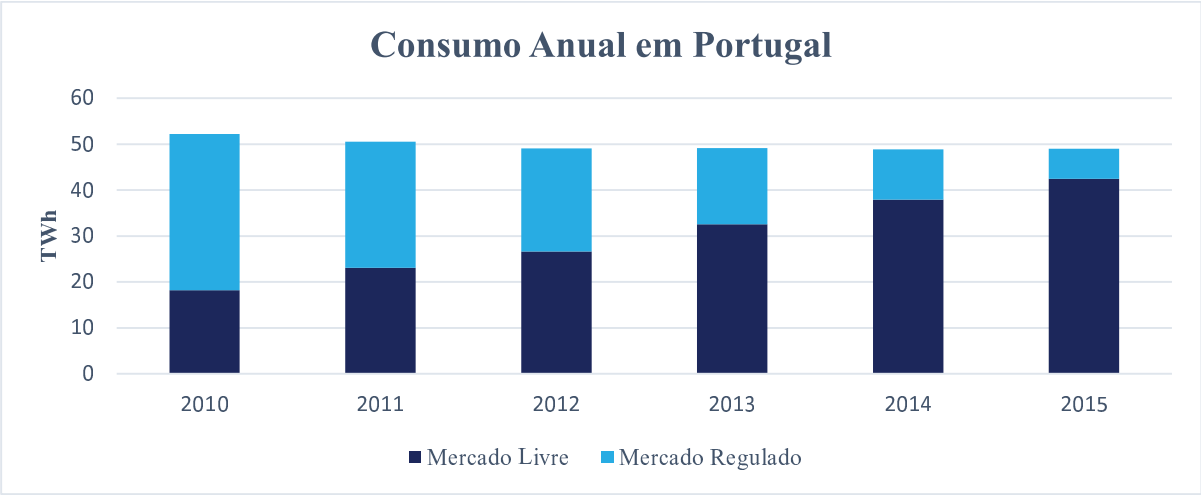

Gráfico 2.5: Evolução do Consumo de Energia Elétrica em Mercado Livre e Regulado desde 2010 a 2015, em Portugal

2.2 Consumo em Portugal

O consumo de energia elétrica varia em função da potência dos aparelhos instalados, do número de horas de utilização dos aparelhos e do número de pessoas que os utilizam. Deste modo, o consumo nacional diário é a soma dos consumos diários de cada cliente de todas as comercializadoras existentes em Portugal Continental.

A cada cliente está associado pelo menos um CPE (Código de Ponto de Entrega), um nível de tensão, uma potência contratada e um contador.

O CPE é um código que identifica a instalação elétrica, ou seja, é um código que reconhece o ponto de entrega da energia. Este código é sempre o mesmo, seja qual for o comercializador.

A potência contratada depende não só do número e da potência dos equipamentos existentes na instalação elétrica, mas também da forma como são utilizados. Quanto maior a quantidade de aparelhos a funcionar em simultâneo, maior deverá ser a potência a contratar.

Deste modo, o nível de tensão associado a cada cliente está dependente da potência contratada para cada instalação elétrica. Existem 4 níveis de tensão: MAT, AT, MT e BT. Este último é definido pela agregação da Baixa Tensão Especial (BTE), Baixa Tensão Normal (BTN) e Iluminação Pública (IP). A Baixa Tensão Normal tem ainda a particularidade de ser segmentada em 3 perfis, A, B e C.

As caraterísticas de cada nível de tensão bem como dos seus perfis relativamente à tensão entre as fases e a potência contratada encontram-se descritas na tabela 2.2.

[6] Fonte: Relatório e Contas 2015, REN

Tabela 2.2: Descrição da Tensão entre Fases e da Potência Contratada por nível de Tensão

Nível de Tensão	Tensão entre Fases	Potência Contratada
MAT	>110kV	≥6 MW
AT	45kV – 110kV	
MT	1kV – 45kV	
BTE	≤ 1kV	> 41,4 kW
BTN		≤ 41,4 kW
A		> 13,8 kVA
B		≤ 13,8 kVA
C		

Analisando a tabela 2.2, o perfil B e o perfil C da BTN contém a mesma potência contratada. O que difere estes perfis é o consumo anual do cliente em si, ou seja, clientes com consumo anual superior a 7140 kWh define o perfil B, enquanto um cliente com consumo anual não superior a 7140 kWh define o perfil C.

A MAT está associada sobretudo as auxiliares de produção de energia elétrica e a indústria siderúrgica, enquanto a AT está relacionada a grandes hospitais, às indústrias de plásticos, às indústrias de celulose, às indústrias de adubos e a serviços energéticos. Por outro lado, a MT está ligada à indústria de componentes automóveis, metalúrgica, moldes, vitrificação, grande hotelaria, entre outros. À BTE e BTN estão associadas sobretudo os clientes residenciais, as lojas, os escritórios e pequenas empresas, enquanto a IP é designada tal como o próprio nome indica à iluminação pública.

É através do contador que se tem acesso ao consumo de energia elétrica em cada instalação. A leitura do contador depende do nível de tensão associado ao cliente. Para os níveis de tensão MAT, AT, MT e para uma parte da BTE, a leitura do contador é feita de modo remoto, isto é, os clientes dispõem de equipamento de medição com registo de consumo em períodos de 15 minutos. Estes níveis de tensão são definidos como sendo telecontados. Para os restantes clientes, a leitura do contador é feita porta a porta por um elemento da distribuidora ou transmitida pelo cliente, via telefone, web, entre outros.

2.2.1 Previsão de Consumos

Para ir de encontro à procura de energia elétrica dos seus clientes, cada comercializadora tem o direito de recorrer ao mercado diário e fazer uma oferta de compra dos consumos de energia elétrica da sua carteira de clientes. É de salientar que a oferta em mercado é realizada na ótica da produção, isto é, tem em conta o consumo para fazer face às necessidades dos clientes e as perdas existentes na rede de distribuição e de transporte.

Como os consumidores não fornecem o volume de energia pretendida, surge, deste modo, a necessidade de realizar previsões de consumos de energia elétrica.

As previsões de consumo de energia elétrica são feitas hora a hora e realizadas todos os dias, uma vez que a oferta de compra dos volumes de energia em mercado é realizada diariamente.

Como anteriormente foi referido, o tempo de disponibilização dos consumos varia consoante o nível de tensão. Enquanto MAT, AT, MT e já uma parte da BTE são telecontados, a obtenção dos consumos reais da BTN depende da interação do cliente ao fornecer a contagem dos seus consumos quer por leitura local, quer por divulgação via telefone/web. A não disponibilização da leitura dos consumos tem

impacto tanto na faturação do cliente, como na do comercializador. O impacto para o consumidor deve-se ao facto de ser faturado um consumo estimado/previsto e não um consumo real. Como durante algum tempo, os consumos da BTN são uma incógnita, surgiu a necessidade de criar um perfil de consumo para este nível de tensão. Anualmente, a ERSE disponibiliza o perfil de consumo de 15 em 15 minutos para cada um dos três perfis da BTN. Deste modo, o impacto para as comercializadoras deve-se ao facto de realizar ofertas de compras de consumos para BTN com base em perfis teóricos.

2.2.2 Diferentes tipos de Consumos

Ao longo da cadeia de valor do setor elétrico surge a necessidade de diferenciar três tipos de consumos: consumo reconciliado, consumo previsto e o consumo real.

A produção de energia elétrica é feita tendo em conta a oferta de mercado de cada comercializador, isto é, o volume de energia a produzir é a soma de todas as necessidades dos clientes de cada comercializadora existente em Portugal. Uma vez que existem diferenças entre o total de energia elétrica que entra na RESP, e a soma dos volumes afetos aos vários comercializadores, torna-se necessário que estas diferenças sejam repartidas de forma proporcional à energia afeta a cada comercializador. A existência desta diferença deve-se à utilização de elementos que não são conhecidos de forma rigorosa, como por exemplo estimativas, perfis de consumo e fatores de ajustamento nos consumos dos seus clientes. Deste modo, o consumo reconciliado para cada comercializador é obtido através da soma da curva telecontada ajustada para perdas com a afetação da curva perfilada ajustada para perdas. A figura 2.2 representa os cálculos efetuados para obter o consumo reconciliado para cada comercializador.

Tal como o próprio nome indica, o consumo previsto é o consumo que advém da elaboração de previsões com o intuito de fazer uma oferta de compra de energia em mercado. Uma vez que as ofertas em mercado são feitas na ótica da produção, o consumo previsto é um consumo com perdas, ou seja, o consumo previsto é uma combinação do consumo necessário para fazer face às necessidades dos clientes e as perdas existentes na rede até chegar ao cliente final.

O consumo real é a soma do volume de energia efetivamente consumido pelos clientes telecontados e o consumo perfilado dos clientes não telecontados. Apesar de ser denominado consumo real, este consumo é em certa parte um consumo teórico, uma vez que para clientes não telecontados é utilizado um perfil de consumo.

Por fim, o volume de energia a ser produzido terá que ter em conta os volumes na ótica do consumidor, isto é, satisfazer as necessidades dos clientes de cada comercializadora e as perdas existentes na rede de distribuição e de transporte. Deste modo, obtém-se a equação 2.1:

$$Produção = Consumo + Perdas \tag{2.1}$$

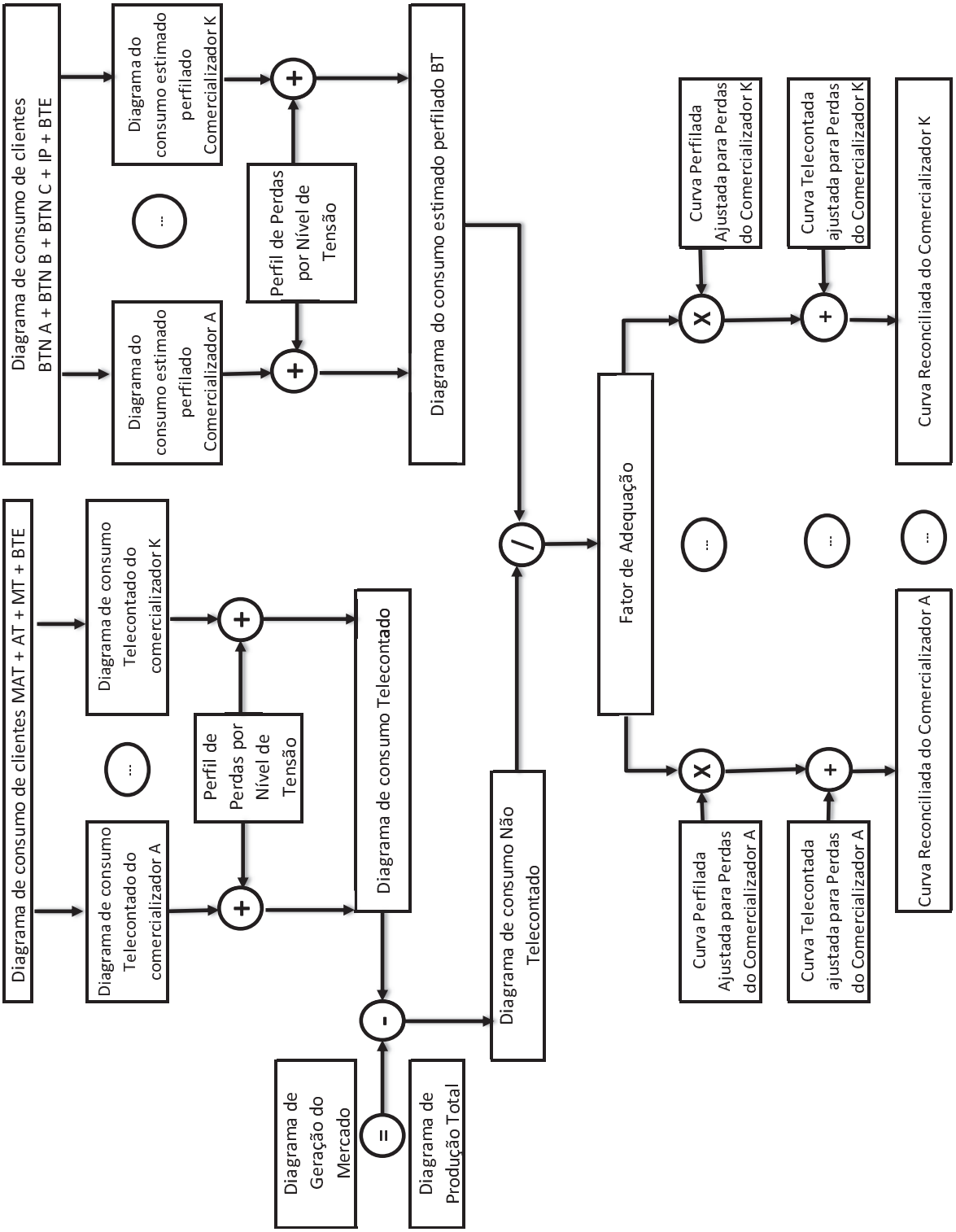

Figura 2.2: Métrica do Consumo Reconciliado[7] - Passo a Passo dos cálculos a efetuar para obter o consumo reconciliado de cada comercializador

[7] Fonte: Guia de Medição, Leitura e Disponibilização de Dados, Janeiro de 2016 - ERSE

Capítulo 3

3. Previsão de Consumos de Energia Elétrica em Portugal Continental

3.1 Introdução

Os modelos de previsão de consumos diários a curto prazo para os quais se têm verificado um melhor desempenho em termos de qualidade da previsão, consideram o consumo num determinado dia *t* como uma combinação linear dos consumos em alguns dias anteriores. Do ponto de vista estatístico, isto significa que o consumo no dia *t*, Z_t, condicional aos consumos em dias anteriores, Z_{t-1}, Z_{t-2}, …, Z_{t-m}, tem valor médio que depende destes últimos e com uma variância pequena. Simbolicamente, a variável aleatória que mede o consumo no dia *t*, dados os valores dos consumos em dias anteriores, isto é,

$$Z_t | Z_{t-1} = Z_{t-1}, \dots, Z_{t-m} = Z_{t-m} \tag{3.1}$$

tem valor médio

$$\mu_t = b_0 + \sum_{j=1}^{m} b_j z_{t-j} \tag{3.2}$$

e variância constante σ^2. Para além deste pressuposto, as previsões de consumo tornam-se mais rigorosas se for possível admitir que a distribuição condicional de Z_t, dadas as *m* observações anteriores, é gaussiana com o valor médio e variância já referidos. Como veremos mais adiante, na secção 3.5, tanto no que respeita à estimação de parâmetros como à inferência estatística, incluindo intervalos de previsão, este modelo pode, até certo ponto, ser tratado como um modelo de regressão múltipla. Assim sendo, começamos por apresentar um resumo sobre o modelo de regressão múltipla: método dos mínimos quadrados, principais propriedades dos estimadores obtidos por este modelo, testes de hipóteses e intervalos de predição.

A análise de regressão é um dos métodos estatísticos mais importantes e utilizados, uma vez que este método conduz a estimadores, testes de hipóteses e intervalos de predição com propriedades estatísticas que, em certo sentido, são ótimas. É a partir de uma função matemática que a regressão linear descreve o comportamento de observações que não dispomos através de outros dados recolhidos e, consequentemente, estuda a relação entre uma variável dependente ou variável resposta e uma ou várias variáveis independentes ou variáveis explicativas. Esta relação é obtida através de uma equação que associa a variável dependente a variáveis independentes.

A relação linear entre a variável dependente e uma variável independente é uma característica de um modelo de regressão linear simples, por outro lado denomina-se modelo de regressão linear múltipla (RLM) a relação entre a variável dependente e várias variáveis independentes.

3.2 O Modelo de Regressão Linear e os seus Pressupostos

Considera-se que a variável dependente Y se pode escrever à custa de um conjunto de m variáveis independentes, $x_1, x_2, \dots, x_m$. Diz-se que um conjunto de observações y_i, $1 \leq i \leq n$, segue um modelo

de regressão linear múltipla se, para as correspondentes observações de variáveis independentes, x_{ij}, j=1, …, m for válida a relação definida na seguinte equação:

$$y_i = \sum_{j=1}^{m} x_{ij} b_j + \varepsilon_i\,, \qquad i = 1, \dots, n \tag{3.3}$$

Os coeficientes $b_1, b_2, \dots, b_m$ são parâmetros a estimar e denominam-se por coeficientes de regressão. Em contrapartida, os ε_i são variáveis aleatórias que verificam os seguintes pressupostos:

1. Os erros ε_i são variáveis aleatórias de média zero, isto é, $E(\varepsilon_i) = 0$.
2. Os erros ε_i são variáveis aleatórias de variância constante, ou seja, $Var(\varepsilon_i) = \sigma^2$.
3. As variáveis aleatórias $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são não correlacionadas, tal que $E(\varepsilon_i \varepsilon_j) = 0$, se i ≠ j.
4. As variáveis aleatórias ε_i são independentes com distribuição normal, N (0, σ^2).

Ao conjunto das três primeiras condições chama-se condições de *Gauss-Markov*. Por vezes, para além destas condições, torna-se necessário exigir que as variáveis aleatórias sejam ainda independentes com distribuição normal de média nula e variância σ^2(Pressuposto 4).

Uma representação equivalente do modelo de regressão linear múltipla descrita na equação 3.3 é:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_m x_{im} + \varepsilon_i, i = 1, \dots, n \tag{3.4}$$

O termo constante b_0 da equação 3.4 corresponde ao coeficiente b_1 da equação 3.3, onde na maior parte dos casos, as variáveis x_{i1} (3.3) são identicamente iguais à unidade, isto é, $x_{i1} = 1, i = 1, \dots, n$.

Um modo mais simples de apresentar o modelo de regressão múltipla é através da notação matricial. Deste modo, as n equações correspondem a uma única equação matricial Y= Xb + ε, em que designamos por:

- Y o vetor de $n \times 1$ das observações da variável dependente;
- X a matriz de planeamento de dimensões $n \times m$ em que a primeira coluna corresponde a um vetor de 1's associada ao termo constante;
- b o vetor de $m \times 1$ de parâmetros do modelo;
- ε o vetor de $n \times 1$ dos n erros aleatórios.

Assim, a representação matricial de cada termo do modelo de regressão linear múltipla encontra-se descrita na equação 3.5.

$$Y = Xb + \varepsilon \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.5}$$

De agora em diante, quando nada é referido, assumir a representação expressa na equação 3.5.

3.2.1 Estimação dos Parâmetros

O método dos mínimos quadrados estima os parâmetros $b_1, b_2, \dots, b_m$ de modo a encontrar o melhor ajustamento para um conjunto de observações no sentido em que minimiza a soma dos quadrados das

diferenças entre o valor estimado e os dados observados. A esta diferença damos o nome de resíduos. Pretende-se através deste método estimar os parâmetros que minimizam a soma dos quadrados dos resíduos. Ou seja, o que se pretende minimizar é a soma dos quadrados como função dos parâmetros tal como se encontra representado na equação 3.6.

$$SQ = SQ(b_1, b_2, \dots, b_m) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} b_j x_{ij})^2 \tag{3.6}$$

Derivando em ordem a cada um dos coeficientes de regressão, obtêm-se:

$$\frac{\partial SQ}{\partial b_k} = 2 \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} b_j x_{ij})(-x_{ik}), k = 1, \dots, m \tag{3.7}$$

Igualando a equação 3.7 a zero, obtém-se o conjunto das equações normais para o modelo de regressão linear múltipla, nomeadamente,

$$\sum_{i=1}^{n} y_i x_{ik} = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} x_{ij} x_{ik} \tag{3.8}$$

Deste modo, trata-se de um sistema de *m* equações lineares a *m* incógnitas. A solução deste sistema é obtida de maneira mais simples utilizando-se notação matricial, ou seja, o modelo de regressão linear pode ser escrito como se encontra representado na equação 3.5.

Em consequência, o sistema de equações normais pode ser reescrito como na equação 3.9. Quando existir a matriz inversa de $X^T X$, o vetor dos estimadores de mínimos quadrados dos coeficientes de regressão é dado pela equação 3.10.

$$(X^T X) b = X^T Y \tag{3.9}$$

$$\hat{b} = (X^T X)^{-1} X^T Y \tag{3.10}$$

Caso a matriz $X^T X$ não seja invertível, isto significa que existe uma ou mais variáveis independentes que são combinação linear de outras e, portanto, essas variáveis devem ser retiradas do modelo.

Finalmente falta verificar se a solução obtida das equações normais corresponde efetivamente a um mínimo da soma dos quadrados. Deste modo, é importante definir os *valores ajustados*. Estes valores são definidos pela equação 3.11:

$$\hat{y}_i = \sum_{j=1}^{m} \hat{b}_j x_{ij}, i = 1, .., n \tag{3.11}$$

Tal como foi referido anteriormente, as estimativas dos termos de erro são os *resíduos,* que são obtidos pelas diferenças entres os valores observados e os valores ajustados, ou seja,

$$e_i = y_i - \hat{y}_i, i = 1, \dots, n \tag{3.12}$$

O que em notação matricial, pode ser escrito como se apresenta na equação 3.13.

$$e = Y - \hat{Y} = Y - X\hat{b} \tag{3.13}$$

Através desta notação é fácil de verificar que, no modelo linear, os resíduos são ortogonais à matriz de planeamento, isto é,

$$X^T e = 0_{m\times 1} \tag{3.14}$$

em que $0_{m\times 1}$ representa o vetor nulo de dimensão $m \times 1$. Deste modo, substituindo o vetor dos resíduos pela sua definição expressa na equação 3.13, isto é equivalente à equação 3.15:

$$X^T e = X^T Y - X^T X\hat{b} = 0_{m\times 1} \tag{3.15}$$

Este conjunto de equações normais, por definição, têm a solução dada pelos estimadores de mínimos quadrados, $\hat{b}$. Como todas as colunas da matriz de planeamento são ortogonais ao vetor de resíduos, se existir uma coluna com todos elementos iguais à unidade, conclui-se que a soma dos resíduos é nula. Outra consequência da ortogonalidade entre as colunas da matriz de planeamento e o vetor de resíduos é a ortogonalidade entre estes e o vetor dos valores ajustados. A demonstração desta consequência encontra-se descrita na equação 3.16.

$$\hat{Y}^T e = \hat{b}^T X^T e = \hat{b}^T 0_{m\times 1} = 0 \tag{3.16}$$

Por último, através destas propriedades, pode-se demonstrar que os EMQ (Estimadores de Mínimos Quadrados) correspondem a um mínimo da soma dos quadrados (SQ). Para ver que assim é, note-se que:

$$\begin{aligned} SQ = (Y - Xb)^T (Y - Xb) &= \left(Y - X\hat{b} + X\hat{b} - Xb\right)^T \left(Y - X\hat{b} + X\hat{b} - Xb\right) \\ &= \left(Y - X\hat{b}\right)^T \left(Y - X\hat{b}\right) + 2\left(\hat{b} - b\right)^T X^T \left(Y - X\hat{b}\right) + \left(\hat{b} - b\right)^T (X^T X)(\hat{b} - b) \end{aligned} \tag{3.17}$$

Como os resíduos são ortogonais à matriz de planeamento vem que:

$$\left(\hat{b} - b\right)^T X^T \left(Y - X\hat{b}\right) = \left(Y - X\hat{b}\right)^T X\left(\hat{b} - b\right) = e^T X\left(\hat{b} - b\right) = 0 \tag{3.18}$$

Assim, simplificando a equação 3.17 com o resultado obtido na equação 3.18, tem-se que a soma dos quadrados é dada por:

$$SQ = \left(Y - X\hat{b}\right)^T \left(Y - X\hat{b}\right) + \left(\hat{b} - b\right)^T (X^T X)\left(\hat{b} - b\right) \tag{3.19}$$

Os termos da soma no lado direito da igualdade são não negativos, uma vez que são somas de quadrados. Por outro lado, uma vez que o primeiro termo desta soma não depende de b, sai que o mínimo será atingido no ponto que anular o segundo termo da soma. Deste modo, o mínimo será atingido em $b = \hat{b}$.

Atente-se ainda que o vetor dos valores ajustados pode ainda ser escrito como função linear do vetor dos valores observados, isto é,

$$\hat{Y} = X\hat{b} = X(X^T X)^{-1} X^T Y = HY, \tag{3.20}$$

em que H é uma matriz $n \times n$, denominada por *hat matrix*, tal que $H = X(X^T X)^{-1} X^T$. Esta matriz é simétrica e idempotente, isto é, $HH = H^2 = H$.

Seja $M = I_n - H$, em que I_n define a matriz identidade $n \times n$, onde esta matriz é também simétrica e idempotente. Além disso, vem que $MX = (I_n - H)X = X - X = 0_{n\times m}$, em que $0_{n\times m}$ define uma matriz $n \times m$ com todos os elementos nulos. Deste modo, estabelece-se uma relação entre o vetor dos resíduos (e) e o vetor de erros aleatórios (ε) dada por:

$$e = Y - \hat{Y} = MY = M(Xb + \varepsilon) = 0_{n\times m}b + M\varepsilon = M\varepsilon \tag{3.21}$$

A relação obtida na equação 3.21 e as condições de *Gauss-Markov* permitem calcular de modo imediato o valor médio e a matriz de covariâncias do vetor dos resíduos. Pela condição 1 de *Gauss-Markov* vem que

$$E(e) = ME(e) = M0_{n\times 1} = 0_{n\times 1} \tag{3.22}$$

e, pelo conjunto das três condições, tem-se que

$$Cov(e) = E(ee^T) = E(M\varepsilon\varepsilon^T M) = M\sigma^2 I_n M = \sigma^2 M = \sigma^2(I_n - H). \tag{3.23}$$

Deste modo, a variância de cada um dos resíduos do modelo de regressão linear múltipla é dada por:

$$Var(e_i) = \sigma^2 m_{ii} = \sigma^2(1 - h_{ii}) \tag{3.24}$$

A condição 1 de *Gauss-Markov* permite ainda concluir que

$$E(Y) = E(Xb + \varepsilon) = Xb, \tag{3.25}$$

e das condições 1 a 3 sai que

$$Cov(Y) = E[(Y - Xb)(Y - Xb)^T] = E(\varepsilon\varepsilon^T) = \sigma^2 I_n. \tag{3.26}$$

3.2.2 Propriedades estatísticas dos EMQ

Nesta secção pretende-se demonstrar as propriedades estatísticas que os EMQ gozam, sob a validade das condições *Gauss-Markov*.

A condição 1 de *Gauss-Markov* garante que o valor esperado dos termos de erro do modelo de regressão linear é nulo, demonstrado na equação 3.22. Esta propriedade em conjunto com a linearidade do valor médio garante que os Estimadores dos Mínimos Quadrados são centrados e tem-se que:

$$E(\hat{b}) = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TE[Y] = (X^TX)^{-1}X^TXb = b \tag{3.27}$$

Se se exigirmos, para além da condição 1, as condições 2 e 3 de *Gauss-Markov*, isto é, se se admitirmos que os termos de erro são não correlacionados e com variância σ^2, resulta que a matriz de covariâncias dos EMQ é definida pela equação 3.28.

$$\begin{aligned} Cov(\hat{b}) &= E[(\hat{b} - b)(\hat{b} - b)^T] \\ &= E\{[(X^TX)^{-1}X^TY - (X^TX)^{-1}X^TE(Y)][(X^TX)^{-1}X^TY - (X^TX)^{-1}X^TE(Y)]^T\} \\ &= E[(X^TX)^{-1}X^T(Y - E(Y))(Y - E(Y))^TX(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \end{aligned} \tag{3.28}$$

A consistência dos EMQ é garantida desde que a soma dos elementos na diagonal da matriz $(X^TX)^{-1}$, tenda para zero quando n→ +∞, ou seja,

$$tr(X^TX)^{-1} \to 0, para\ n \to +\infty \tag{3.29}$$

Quando esta condição é satisfeita, verifica-se que cada um dos m elementos da diagonal de $(X^TX)^{-1}$ converge para zero quando n→ +∞. Assim, representando por $z_{jj}, j = 1, \dots, m$, os elementos da diagonal da matriz $(X^TX)^{-1}$, tem-se que $z_{jj} \to 0$, quando $n \to +\infty$. Sai também que:

$$Var(\hat{b}_j) = \sigma^2 z_{jj} \to 0 \tag{3.30}$$

Finalmente, através da propriedade dos EMQ serem centrados (3.27) e da equação (3.30) pode-se concluir que $\hat{b}_j$converge em probabilidade para b_j, isto é, $\hat{b}_j \xrightarrow{p} b_j$.

O valor médio e a matriz de covariâncias do vetor dos valores ajustados são definidos por:

$$E(\hat{Y}) = Xb\,, \tag{3.31}$$

$$Cov(\hat{Y}) = E[X(\hat{b} - b)(\hat{b} - b)^T X^T] = XCov(\hat{b})X^T = \sigma^2 X(X^T X)^{-1} X^T = \sigma^2 H, \tag{3.32}$$

em que H é denominada por *hat matrix* e a $Var(\hat{y}_i) = \sigma^2 h_{ii}$.

3.2.3 Coeficientes de Regressão: Testes e Intervalos de Confiança

Construído o modelo de regressão linear múltipla, torna-se necessário testar a nulidade de alguns dos parâmetros b_j, ou seja, é realizado o teste $\boldsymbol{H_0: b_j = 0}$ **contra** $\boldsymbol{H_1: b_j \neq 0}$, para j=1, …, *m*. A importância deste teste deve-se ao facto de se estar a testar se a variável explicativa x_j influencia ou não a variável dependente Y e, consequentemente, se deverá ser ou não incluída no modelo de regressão linear múltipla.

Deste modo, antes de se iniciar o teste acima, é necessário ter em consideração os resultados descritos no teorema 3.1.

Teorema 3.1. Seja Y = Xb + ε um modelo linear em que ε é um vetor de variáveis aleatórias i.i.d. com distribuição normal N(0, σ^2). Então:

1. O EMQ do vetor de parâmetros **b**, isto é, $\hat{b} = (X^T X)^{-1} X^T Y$ tem distribuição multinormal, N (b,$\sigma^2 (X^T X)^{-1}$).
2. A variável aleatória $\frac{(n-p)S^2}{\sigma^2} = \frac{e^T e}{\sigma^2}$ tem distribuição qui-quadrado com *n-m* graus de liberdade.
3. $\hat{b}$ e S^2 são independentes.

A variância dos parâmetros b_j é dada por:

$$Var(b_j) = \sigma^2 z_{jj} \tag{3.33}$$

em que z_{jj} é o j-ésimo elemento da diagonal da matriz $(X^T X)^{-1}$. Deste modo, a estatística de teste pode ser escrita na forma:

$$\frac{\hat{b}_j}{S\sqrt{z_{jj}}} = \frac{\hat{b}_j}{\sigma(\hat{b}_j)} \tag{3.34}$$

onde, sob a validade de H_0, tem distribuição t de student com *n-m* graus de liberdade.

A região de rejeição deste teste é dada por:

$$\frac{|\hat{b}_j|}{\sigma(\hat{b}_j)} > t_{n-m}^{1-\alpha/2}$$

(3.35)

3.2.4 Intervalos de Predição

Considere-se um conjunto de valores não observados de variáveis independentes e seja X^* a matriz com a representação desses valores definida na equação 3.36.

$$X^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_m^* \end{bmatrix} \tag{3.36}$$

Seja Y^* a variável aleatória dada pela equação 3.37, em que ε^* tem distribuição normal de média nula e variância σ^2.

$$y^* = X^* b + \varepsilon^* = \sum_{j=1}^{m} x_j^* b_j + \varepsilon^*$$

(3.37)

O que se pretende nesta secção é um intervalo de confiança para a variável dependente y^*. Deste modo, a predição de $\hat{y}^*$ é dada por:

$$\hat{y}^* = \sum_{j=1}^{m} x_j^* \hat{b}_j = X^{*T} \hat{b}$$

(3.38)

Tem-se que $E(\hat{y}^*) = E(y^*)$ e o erro quadrático médio (EQM) de $\hat{y}^*$ é dado por:

$$\begin{aligned} E[(\hat{y}^* - y^*)^2] &= E[(\hat{y}^* - y^*)(\hat{y}^* - y^*)^T] = E\left[\left(X^{*T}\hat{b} - X^{*T}b - \varepsilon^*\right)\left(X^{*T}\hat{b} - X^{*T}b - \varepsilon^*\right)^T\right] \\ &= E\left[X^{*T}\left(\hat{b} - b\right)\left(\hat{b} - b\right)^T X^*\right] + E\left(\varepsilon^* \varepsilon^{*T}\right) = \sigma^2[X^{*T}(X^T X)^{-1}X^* + 1] \end{aligned}$$

(3.39)

Veja-se ainda que o erro de predição $\hat{y}^* - y^* = X^{*T}(\hat{b} - b) + \varepsilon^*$ é uma combinação linear do vetor dos EMQ, que tem distribuição normal multivariada, e do termo de erro ε^*, o qual tem também distribuição normal e independente do vetor de estimadores. Então o erro de predição tem distribuição normal com média nula e variância igual ao EQM de $\hat{y}^*$. Deste modo, a variável

$$\frac{\hat{y}^* - y^*}{S\sqrt{X^{*T}(X^T X)^{-1}X^* + 1}}$$

(3.40)

tem distribuição *t-student* com *n-m* graus de liberdade, a partir da qual se constrói um intervalo de $(1-\propto) \times 100\%$ de confiança para y^*. Assim um intervalo de predição para a variável dependente é definido por

$$\left(\hat{y}^* - t_{n-m}^{1-\frac{\propto}{2}} S\sqrt{X^{*T}\left(X^T X\right)^{-1} X^* + 1};\ \hat{y}^* + t_{n-m}^{1-\frac{\propto}{2}} S\sqrt{X^{*T}\left(X^T X\right)^{-1} X^* + 1}\right) \tag{3.41}$$

em que $t_{n-m}^{1-\frac{\propto}{2}}$ representa o quantil de ordem $1-\frac{\propto}{2}$ da distribuição *t-student* com *n-m* graus de liberdade.

3.3 Validação do Modelo

Criado o modelo de regressão linear múltiplo é necessário validar alguns pressupostos que garantem a qualidade das inferências e predições produzidas pela análise de regressão. Alguns dos pressupostos a validar são as condições de *Gauss-Markov* e a normalidade dos resíduos.

Nesta secção dar-se-á destaque à análise dos resíduos e ao coeficiente de determinação.

3.3.1 Análise dos Resíduos

Como já foi referido anteriormente, os resíduos são definidos pela diferença entre os valores observados e os valores ajustados como se pode verificar através da equação 3.12. Deste modo, os resíduos demonstram as disparidades entre a realidade e o modelo construído.

A análise dos resíduos do modelo de regressão linear é constituída pela:

1. Representação gráfica dos resíduos contra cada uma das variáveis independentes incluídas no modelo de regressão linear múltiplo.
2. Representação gráfica dos resíduos contra outras variáveis independentes que não tenham sido incluídas no modelo.
3. Representação gráfica dos resíduos contra os valores ajustados.
4. Estudo da normalidade dos resíduos através de histogramas e testes de ajustamento.

A análise descrita em 1 permite verificar se faz ou não sentido incluir no modelo alguma das variáveis independentes utilizadas ou, em alternativa, se há necessidade de esta sofrer alguma transformação antes de ser incluída no modelo de regressão linear múltipla.

Através da representação descrita no ponto 2 podemos detetar a existência de alguma relação entre a variável dependente e essas novas variáveis e, nesse caso, a variável em análise deve ser incluída no modelo.

A representação gráfica em 3 ajuda a identificar a necessidade de modificar ou juntar novas variáveis.

A realização de histogramas e testes de ajustamento, tais como o teste do Qui-Quadrado, teste de *Lilliefors* ou o teste de *Kolmogorov-Smirnov* no ponto 4 permitem verificar a normalidade dos resíduos. Estes testes são, frequentemente, realizados com os resíduos padronizados, ou seja, divididos pelo seu desvio padrão e são dados por:

$$e_i^* = \frac{e_i - \bar{e}}{S} \tag{3.42}$$

O teste usado para o estudo da normalidade dos resíduos foi o teste de *Lilliefors*, em que as hipóteses de teste são:

H_0: Os resíduos provém de uma distribuição normal *versus* **H_1**: Os Resíduos não provém de uma distribuição normal.

Deste modo, torna-se necessário perceber como é realizada a construção do teste de *Lilliefors* para a normalidade dos resíduos. Assim, denote-se:

1. $S(x_i) = \frac{i}{n}$, para *i*=1,…,*n,* em *n* que define o tamanho da amostra.

2. $Z_i = \frac{e_i - \bar{e}}{S}$, para *i*=1,…,*n*, em que e_i representa os resíduos, $\bar{e}$ e *S* representam, respetivamente, a média e o desvio padrão dos resíduos do modelo de regressão múltipla.

Os resíduos seguem uma distribuição normal no caso em que *d* é inferior a $dc_{n,\propto}$, sendo $dc_{n,\propto}$,[8] um valor tabelado para o teste de *Lilliefors* com tamanho *n*, ao nível de significância α, e *d* é definido pelo máximo entre o módulo da diferença da f.d.p. de Z_i e $S(x_{i-1})$ e o módulo da diferença da f.d.p. de Z_i e $S(x_i)$, isto é,

$$d = \max_{1 \le i \le n} \{|F(Z_i) - S(x_{i-1})|, |F(Z_i) - S(x_i)|\} \tag{3.43}$$

Assim, caso $d \geq dc_{n,\propto}$, a hipótese nula é rejeitada e, deste modo, os resíduos não seguem uma distribuição normal.

3.3.2 Coeficiente de determinação R^2

Considere-se o modelo de regressão com um termo constante, ou seja, a variável x_{i1} é igual à unidade para $i = 1, \dots, n$. Seja SQ_{Tot} a variabilidade total da amostra definida pela equação 3.44, onde $\bar{y}$ representa a média da amostra. Tem-se a seguinte decomposição

$$SQ_{Tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SQ_e + SQ_{Reg} \tag{3.44}$$

A parcela SQ_e representa a soma dos quadrados dos resíduos e é dada por:

$$SQ_e = \sum_{i=1}^{n} e_i^2 \tag{3.45}$$

Por outro lado, a parcela SQ_{Reg} descreve a soma dos quadrados do modelo de regressão e é definida por:

$$SQ_{Reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.46}$$

O coeficiente de determinação múltipla, R^2, define a percentagem de variação da amostra que é explicada pelo modelo de regressão, isto é,

$$R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_e}{SQ_{Tot}} \tag{3.47}$$

[8] Visualizar Anexo 1 com a tabela de Estatística de teste de *Lilliefors* para a distribuição Normal

Este coeficiente está compreendido entre 0 e 1. No caso em que o coeficiente de determinação múltipla toma valores iguais à unidade corresponde a um modelo com ajustamento perfeito, ou seja, os valores ajustados correspondem exatamente aos valores observados e, consequentemente, os resíduos do modelo de regressão são nulos. Para o caso em que R^2 é nulo, a variabilidade total da amostra é igual à soma dos quadrados dos resíduos, ou seja, a variabilidade da amostra resulta apenas da variabilidade dos erros do modelo de regressão linear múltipla e as variáveis independentes em nada contribuem para a explicação dessa variabilidade.

Se o coeficiente de determinação múltipla for apresentado de outro modo, como se poderá visualizar na equação 3.48, pode-se ver que é equivalente ao coeficiente de correlação amostral entre os valores observados e os valores ajustados, isto é, mede o grau de associação linear entre a amostra observada e os valores ajustados.

$$R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{[\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \qquad = \left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}}\right]^2 = r_{y\hat{y}}^2$$

(3.48)

É de salientar que para valores pequenos da dimensão da amostra, o valor do coeficiente de determinação múltipla tende a estar inflacionado. Deste modo, a análise do R^2 é realizada sempre em conjunto com o estudo da variância da amostra, S^2. Assim, o estudo de S^2 passa por verificar se a amplitude do intervalo de predição com 95% de confiança para amostras de dimensão grande é aproximadamente 4S.

É de notar que o coeficiente de determinação múltipla para modelos sem termo constante é definido de outro modo, como mostra a equação 3.49.

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

(3.49)

3.4 Deteção de Multicolinearidade

Nesta secção irá ser validada a qualidade dos estimadores obtidos pelo método dos mínimos quadrados. Anteriormente foi referido que o método dos mínimos quadrados, sob as condições de *Gauss-Markov*, produz bons estimadores para os coeficientes de regressão. Porém, a grandeza das suas variâncias pode ser bastante afetada se existirem variáveis independentes que estejam relacionadas entre si, provocando assim um mau ajustamento do modelo.

Suponha-se, por exemplo, a existência de uma relação linear entre duas variáveis x_l e x_k, $x_l = a x_k$. Desta forma, a estimação pelos mínimos quadrados pode produzir quaisquer valores para os coeficientes dessas variáveis, b_l e b_k, desde que $b_l + a\, b_k$ esteja muito próximo da real contribuição da variável x_k para a variabilidade de y. A consideração de variáveis relacionadas entre si no modelo, poderá provocar um aumento na variância dos estimadores de ambos os coeficientes, o que conduzirá à não rejeição de cada uma ou de ambas as hipóteses de que esses coeficientes poderão ser nulos. A estas situações dá-se

o nome de multicolinearidade. Nesta secção explicar-se-á em detalhe o conceito de multicolinearidade e formas de o evitar em modelos de regressão linear.

Seja X a matriz de planeamento de um modelo de regressão linear múltipla tal que $X = [x_{[1]} \quad x_{[2]} \quad x_{[3]} \quad \dots \quad x_{[m]}]$, em que $x_{[j]}$ representa a coluna j dessa matriz, ou seja, representa o conjunto das observações da j-ésima variável independente, $x_{ij}, i = 1, \dots, n$. Caso existam colunas da matriz de planeamento linearmente dependentes, então $X'X$ é uma matriz singular e os estimadores de mínimos quadrados do vetor de coeficientes b não são únicos. Uma solução para este problema será retirar uma das variáveis independentes que seja combinação linear de outras. Uma situação mais complexa será quando existe um subconjunto de colunas X que são apenas aproximadamente linearmente dependentes.

Deste modo, cria-se a necessidade de explicar o conceito de quase singularidade. A singularidade pode ser definida em termos da existência de um vetor c unitário, em que $c' = [c_1 \quad \dots \quad c_m]$ e $cc' = 1$, tal que $Xc = 0$ ou $c'X'Xc = 0$. Pode-se também definir quase singularidade em termos da existência de um vetor c unitário tal que $\|Xc\|^2 = c'X'Xc = \delta$, em que δ está próximo de zero, ao que é equivalente dizer que, para algum vetor unitário c, a norma de $\sum_{j=1}^{m} c_j X_{[j]}$ está próxima de zero.

Quando a quase singularidade se verifica nas colunas da matriz X, a variância dos EMQ pode ser bastante inflacionada. Com isto, os resultados de testes t ou F deixam de ser fiáveis, uma vez que o aumento da variância diminui o valor da estatística de teste e conduz à não rejeição de H_0, mesmo quando as variáveis do modelo de regressão linear múltipla sejam significativas. Assim,

$$1 = (c'c)^2 = \left[c'(X'X)^{1/2}(X'X)^{-1/2}c\right]^2 \tag{3.50}$$

Pela desigualdade de *Cauchy-Schwarz*, tem-se que:

$$1 \leq c'(X'X)cc'(X'X)^{-1}c = \delta c'(X'X)^{-1}c \tag{3.51}$$

Como $Var(b) = \sigma^2(X'X)^{-1}$, a variância de $c'b$ é definida pela equação 3.52, em que será grande quando δ for pequeno em comparação com σ^2.

$$Var(c'b) = \sigma^2 c'(X'X)^{-1}c \geq \frac{\sigma^2}{\delta} \tag{3.52}$$

Deste modo, resulta que alguns dos estimadores dos coeficientes, $\hat{b}_j$, têm variâncias grandes. Assim, a multicolinearidade é o caso específico da quase singularidade, uma vez que existe uma relação linear aproximada entre duas ou mais colunas $X_{[j]}$'s. Dito de outra maneira, a norma do vetor $\sum_{j=1}^{m} c_j X_{[j]}$ é pequena com, pelo menos, dois dos $X_{[j]}$'s e correspondentes coeficientes não tão pequenos. Uma vez que a combinação linear do vetor anterior é afetada pelas unidades em que são medidas as distintas variáveis independentes, é necessário reduzir à mesma escala as variáveis independentes.

Deste modo, em vez de se considerar o modelo $Y = Xb + \varepsilon$, considerar-se-á um modelo equivalente definido na equação 3.53, em que $X_{(s)} = XD_{(s)}^{-1}$ e $b_{(s)} = D_{(s)}b$, com $D_{(s)} = diag(\|X_{[1]}\|, \|X_{[2]}\|, \dots, \|X_{[m]}\|)$.

$$Y = X_{(s)}b_{(s)} + \varepsilon \tag{3.53}$$

Seja $\hat{b}_{(s)}$ o estimador de mínimos quadrados de $b_{(s)}$, assim $\hat{b}_{(s)} = D_{(s)}\hat{b}$ com $Cov(\hat{b}_{(s)}) = D_{(s)}Cov(\hat{b})D_{(s)}$.

3.4.1 Tolerâncias e fatores de inflação das variâncias

O grau de dependência entre cada variável independente x_j e as restantes variáveis incluídas no modelo de regressão linear múltipla é avaliado através da análise dos valores de $R_j^2, j = 1, \dots, m$, em que R_j^2 representa o valor de R^2 quando se executa uma regressão de x_j sobre o conjunto das restantes variáveis. Estuda-se assim o valor de R^2 para o modelo representado na equação 3.54.

$$x_j = a_1 x_1 + \cdots + a_{j-1} x_{j-1} + a_{j+1} x_{j+1} + \cdots + a_p x_p + \eta \tag{3.54}$$

A tolerância da variável independente x_j, designada por TOL_j, é definida por:

$$TOL_j = 1 - R_j^2 \tag{3.55}$$

Quando o valor de TOL_j se encontra muito próximo da unidade, é equivalente a dizer que a variável x_j é independente das restantes. Caso contrário, se TOL_j estiver próxima de zero, significa a existência de uma relação aproximadamente linear entre x_j e alguma das restantes variáveis independentes do modelo de regressão linear múltipla.

O estudo de multicolinearidade do modelo é também efetuado através da análise do fator de inflação da variância. Este fator é representado por VIF_j e corresponde ao inverso da tolerância da variável independente x_j.

$$VIF_j = \frac{1}{TOL_j} \tag{3.56}$$

É imediata a interpretação dos resultados obtidos pela equação 3.56. Um valor de VIF_j muito próximo da unidade indica que não há dependência entre a variável x_j e as restantes variáveis incluídas no modelo, enquanto valores grandes indicam a presença de multicolinearidade.

Uma forma de obter os fatores de inflação da variância é através da análise da diagonal de R^{-1}, em que R representa a matriz de correlações das variáveis independentes do modelo de regressão linear múltipla. Os elementos na diagonal de R^{-1} são exatamente os fatores de inflação da variância, isto é, se $R^{-1} = [r_{jj}]$, então tem-se que $r_{jj} = VIF_j$. Deste modo, aconselha-se a retirar as variáveis correspondentes a entradas elevadas nesta diagonal. Quando existem diversas variáveis nestas condições deve-se retirar a variável correspondente ao elemento da diagonal mais elevado e, em seguida, recalcular a inversa da matriz de correlações e repetir o processo até que não exista nenhum elemento na diagonal demasiado elevado. O objetivo deste processo é retirar as variáveis que causem multicolinearidade.

3.4.2 Valores Próprios e Números Condição

Considere-se a matriz $X'_{(s)}X_{(s)}$, em que a matriz $X_{(s)}$ foi definida anteriormente. É de salientar que a soma dos valores próprios de uma matriz é igual ao seu traço e ainda que cada elemento na diagonal da matriz $X'_{(s)}X_{(s)}$ é unitário. Deste modo, tem-se a seguinte igualdade:

$$\sum_{j=1}^{m} \lambda_j = tr\left(X'_{(s)}X_{(s)}\right) = m \tag{3.57}$$

Os valores próprios de $X'_{(s)}X_{(s)}$ são representados por λ_j, para j=1, …, m. Os valores próprios próximos de zero indicam a existência de dependência linear entre colunas da matriz de planeamento.

A grandeza de um valor próprio relativamente a outros pode ser avaliada através dos números de condição, η_j, que são definidos por

$$\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}} \tag{3.58}$$

em que $\lambda_{max} = \max_{1 \le j \le m} \lambda_j$. Assim, um valor próprio que contenha um número de condição não inferior a 30 indica a existência de uma relação linear entre variáveis da matriz de planeamento, devendo, deste modo, identificar-se as variáveis intervenientes nessa relação, quer através da análise da diagonal da inversa da matriz de correlações, quer através da análise das componentes da variância descrita na próxima secção.

3.4.3 Componentes da Variância

Uma maneira de identificar quais as combinações lineares das colunas da matriz de planeamento X que causam multicolinearidade é através da análise da matriz dos vetores próprios de $X'_{(s)}X_{(s)}$ e do cálculo da sua influência na variância dos estimadores dos coeficientes de regressão. Esta análise é realizada utilizando a matriz $X_{(s)}$. Uma vez que $X'_{(s)}X_{(s)}$ representa uma matriz simétrica, pode-se reescrevê-la como $X'_{(s)}X_{(s)} = \Gamma D_\lambda \Gamma'$, em que $D_\lambda = diag(\lambda_1, \dots, \lambda_m)$ e a matriz Γ, descrita pela equação 3.59, é a matriz ortogonal cujas colunas representam os vetores próprios de $X'_{(s)}X_{(s)}$.

$$\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{bmatrix} \tag{3.59}$$

Deste modo, a matriz de covariâncias de $\hat{b}_{(s)} = D_{(s)}\hat{b}$ pode ser escrita como $Cov(\hat{b}_{(s)}) = \sigma^2 (X'_{(s)}X_{(s)})^{-1} = \sigma^2 \Gamma D_\lambda^{-1} \Gamma'$ e a variância de cada elemento do vetor dos EMQ dos coeficientes de regressão é dada pela equação 3.60.

$$var\left(\hat{b}_j^{(s)}\right) = \sigma^2 \sum_{k=1}^{m} \lambda_k^{-1} \gamma_{jk}^2 \tag{3.60}$$

A $\lambda_k^{-1}\gamma_{jk}^2$ dá-se o nome de componentes da variância de $\hat{b}_j^{(s)}$, para k=1, …, m, e os coeficientes

$$\phi_{kj} = \frac{\lambda_k^{-1}\gamma_{jk}^2}{\sum_{l=1}^{m} \lambda_l^{-1}\gamma_{jl}^2} \tag{3.61}$$

correspondem à proporção da variância do j-ésimo coeficiente de regressão, $\hat{b}_j^{(s)}$, que pode ser explicada pelo k-ésimo valor próprio.

Uma vez que a cada valor próprio nulo corresponde apenas uma única relação linear entre variáveis, a análise das componentes da variância para cada um desses valores próprios indica quais são as variáveis envolvidas em cada uma das relações existentes. Deste modo, para um determinado valor próprio

próximo de zero, as componentes da variância próximas da unidade indicam as variáveis independentes envolvidas nessa relação. Logo, poder-se-á identificar as variáveis que provocam um aumento excessivo da variância de alguns estimadores dos coeficientes de regressão, através da construção da tabela 3.1.

Tabela 3.1: Valores Próprios, números de condição e proporções de variância para os estimadores dos coeficientes de regressão

Valores Próprios	Números de Condição	Proporções de $Var(\hat{b}_1)$	...	$Var(\hat{b}_m)$
λ_1	η_1	$\frac{\sigma^2\gamma_{11}^2}{\lambda_1 Var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{m1}^2}{\lambda_1 Var(\hat{b}_m)}$
λ_2	η_2	$\frac{\sigma^2\gamma_{12}^2}{\lambda_2 Var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{m2}^2}{\lambda_2 Var(\hat{b}_m)}$
...	...	...	...	...
λ_m	η_m	$\frac{\sigma^2\gamma_{1m}^2}{\lambda_m Var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{mm}^2}{\lambda_m Var(\hat{b}_m)}$

3.5 Seleção de Variáveis

Como já referimos, apresentamos neste capítulo o modelo de regressão linear múltipla e as suas propriedades com vista à sua utilização na construção de um modelo para as previsões do consumo diário de energia elétrica recorrendo a variáveis independentes que incluem os consumos em dias anteriores. Assim sendo, não se trata exatamente de um modelo de um modelo de regressão, porque o consumo diário, digamos no dia *t*, corresponde à variável dependente nesse dia, mas será utilizado com variável independente no dia *t*+1 e seguintes. O que vamos ver no início desta secção é que podemos ajustar um modelo, na verdade, um processo *Markoviano*, a estas observações de tal modo que, no que respeita à seleção de variáveis, pode ser tratado como um modelo de regressão linear múltipla.

Consideremos então uma sucessão de variáveis aleatórias, $Z_1, Z_2, \ldots, Z_t, \ldots$, ordenadas sequencialmente no tempo, e que segue um processo *markoviano* de ordem *m*, *m* inteiro positivo, isto é,

$$Z_t|Z_{t-1}, Z_{t-2}, \ldots, Z_{t-j} \overset{d}{=} Z_t|Z_{t-1}, Z_{t-2}, \ldots, Z_{t-m} \tag{3.62}$$

qualquer que seja $j=m+1, m+2, \ldots$. Admita-se ainda que, para cada t, $1 \leq t \leq n$, a distribuição condicional de $Z_t|Z_{t-1} = z_{t-1}, \ldots, z_{t-m} = z_{t-m}$ ou, mais abreviadamente, $Z_t|z_{t-1}, \ldots, z_{t-m}$ é Gaussiana com valor médio

$$E[Z_t|z_{t-1}, \ldots, z_{t-m}] = b_0 + \sum_{j=1}^{m} b_j z_{t-j} \tag{3.63}$$

e variância constante, isto é,

$$Var[Z_t|z_{t-1}, \ldots, z_{t-m}] = \sigma^2. \tag{3.64}$$

Para um conjunto de observações, $Z_1, Z_2, \dots, Z_n$, os parâmetros $b_0, b_1, \dots, b_m$, podem ser estimados pelo método da máxima verosimilhança, ou seja, por maximização da verosimilhança condicional. Deste modo, a equação 3.65 representa a verosimilhança condicional do modelo em estudo:

$$\begin{aligned} L(z_{m+1}, z_{m+2}, \dots, z_n | z_1, \dots, z_m) &= \prod_{t=m+1}^{n} f(z_t | z_{t-1}, \dots, z_{t-m}) \\ &= \prod_{t=m+1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} \left(z_t - b_0 - \sum_{j=1}^{m} b_j z_{t-j} \right)^2 \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=m+1}^{n} \left[z_t - b_0 - \sum_{j=1}^{m} b_j z_{t-j} \right]^2 \right) \end{aligned} \tag{3.65}$$

Daqui pode-se concluir que a logverosimilhança é dada por:

$$\ln L = \ln L(z_{m+1}, z_{m+2}, \dots, z_n | z_1, \dots, z_m) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=m+1}^{n} \left(z_t - b_0 - \sum_{j=1}^{m} b_j z_{t-j} \right)^2 \tag{3.66}$$

Deste modo, as derivadas parciais da logverosimilhança em ordem a cada um dos coeficientes são dadas por, respetivamente,

$$\frac{\partial \ln L}{\partial b_0} = \frac{1}{2\sigma^2} \sum_{t=m+1}^{n} \left(z_t - b_0 - \sum_{j=1}^{m} b_j z_{t-j} \right) \tag{3.67}$$

e

$$\frac{\partial \ln L}{\partial b_k} = \frac{1}{2\sigma^2} \sum_{t=m+1}^{n} \left(z_t - b_0 - \sum_{j=1}^{m} b_j z_{t-j} \right) z_{t-k} \tag{3.68}$$

para *k*=1, …, *m.* Assim, as equações normais são definidas pelas derivadas parciais, representadas nas equações 3.67 e 3.68, igualadas a zero, ou seja, as equações normais são exatamente as mesmas do que as que se obtêm para um modelo de regressão linear múltipla, em que se escreve Z_t como um modelo de regressão linear sobre $Z_{t-1}, \dots, Z_{t-m}$. Portanto, os estimadores de máxima verosimilhança têm exatamente a mesma expressão do que os estimadores de mínimos quadrados para um modelo de regressão múltipla em que a matriz de planeamento é dada pela equação 3.69.

$$Z_m = \begin{bmatrix} 1 & Z_m & Z_{m-1} & \cdots & Z_1 \\ 1 & Z_{m+1} & Z_m & \cdots & Z_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n-1} & Z_{n-2} & \cdots & Z_{n-m} \end{bmatrix} \tag{3.69}$$

O vetor de variáveis dependentes é dado por $z' = [Z_{m+1} \quad Z_{m+2} \quad \cdots \quad Z_n]$, ou seja, $\hat{b} = (Z'_m Z_m)^{-1} Z'_m z$. Para este modelo já não é tão simples deduzir as propriedades estatísticas destes estimadores. No entanto, uma vez que a f.d.p condicional de $Z_t | z_{t-1}, \dots, z_{t-m}$ verifica as condições do teorema da máxima verosimilhança, podemos dizer que assintoticamente, têm distribuição normal, são centrados com variância igual ao limite inferior de *Cramer-Rao*.

Uma vez que as condições do teorema de máxima verosimilhança são verificadas, pode-se tratar o modelo de previsão de consumos de energia elétrica como sendo um modelo de regressão linear múltipla condicional a consumos históricos. Assim, os métodos estatísticos explicados ao longo deste capítulo tornam-se válidos para a obtenção de um modelo de previsão de consumos com bons resultados.

Por fim, torna-se necessário estudar quais são as variáveis que melhor explicam o modelo de previsão de consumos. Inicialmente constrói-se um modelo de regressão múltipla com várias variáveis independentes, onde se julga que estas variáveis influenciam a variável dependente. Mas, por outro lado, torna-se vantajoso reduzir o número de variáveis utilizadas no modelo, uma vez que a utilização de muitas variáveis aumenta assim a possibilidade de existência de multicolinearidade no modelo.

O objetivo desta secção passa por construir um modelo parcimonioso, ou seja, ter um modelo de regressão com o mínimo possível de variáveis mantendo sempre a qualidade do ajustamento. Quanto maior for o número de parâmetros a estimar, maior será a variância dos estimadores do modelo e, por isso, dever-se-á evitar a inclusão de variáveis que não tenham um peso significativo na explicação da variável resposta.

É de se esperar que não exista um método que produza melhores resultados do que outros, assim é conveniente utilizar mais do que um método e escolher aquele que faz mais sentido ou que revela um melhor ajustamento. Note-se que métodos diferentes poderão conduzir a resultados diferentes, mas semelhantes a nível de ajustamento. No entanto, torna-se difícil dizer qual dos modelos é o melhor. Por vezes diferentes métodos produzem o mesmo modelo, o que é uma boa indicação da qualidade de ajustamento do modelo produzido.

Os métodos de seleção de variáveis utilizados na construção de vários modelos de teste de previsão de consumos de energia elétrica em Portugal foram o método da seleção regressiva, também conhecido por *Backward Selection*, e o método de seleção *Stepwise*.

3.5.1 Método de Seleção Regressiva

O método da seleção regressiva é um método onde é considerado um modelo com todas as variáveis em estudo e elimina, passo a passo, a variável cujo teste *t* é menos significativo, ou seja, retira a variável com menor influência na variável resposta. Na fase posterior, reajusta-se o modelo com as restantes variáveis e repete-se o processo até que todas as variáveis contidas no modelo sejam significativas.

De uma forma resumida, designemos por *V* o número total de variáveis e por *v* o número de variáveis no modelo.

Passo 1. Iniciar o modelo com todas as *V* variáveis.

Passo 2. Ajustar o modelo com as *V* variáveis e calcular os quocientes *t* para cara uma delas.

Passo 3. Caso o quociente t mais pequeno seja não significativo: eliminar a variável correspondente, ou seja, aquela à qual corresponde maior valor-P[9]. Considerar v=v-1 e retomar o passo 2.

Passo 4. Caso o quociente t mais pequeno seja significativo: nenhuma das v variáveis do modelo pode ser omitida e, deste modo, adotar o modelo atual.

É de salientar que o modelo final depende do nível de significância α escolhido para o teste t e, ainda que, a presença de multicolinearidade forte entre as variáveis faz com que este método produza resultados pouco fiáveis.

3.5.2 Método de Seleção *Stepwise*

O método de seleção *Stepwise* é um método de seleção progressiva. Este método funciona de modo inverso ao método anteriormente apresentado, uma vez que se começa com zero variáveis no modelo e que se vão introduzindo, progressivamente, aquelas que provocam maior aumento na qualidade do ajustamento, ou seja, no R^2 ou, consequentemente, uma maior redução na soma dos quadrados dos resíduos. Este processo finda quando não existe mais nenhuma variável significativa a ser incluída no modelo.

De modo análogo ao que foi apresentado no método de seleção regressiva, consideremos V o número total de variáveis e v o número de variáveis no modelo.

Passo 1. Iniciar o modelo com zero variáveis, ou seja, v=1.

Passo 2. Considerar cada uma das restantes V-v variáveis e escolher aquela que provoca maior redução na soma dos quadrados dos resíduos, SQ_e, ou maior aumento no R^2.

Passo 3. Testar se a redução na SQ_e é significativa, realizando um teste t ou um teste F na variável escolhida.

Passo 4. Se o resultado do teste do passo 3 for negativo juntar a variável ao modelo. Caso contrário ir para o passo 6.

Passo 5. Após a inclusão da nova variável, testar a significância das restantes variáveis inseridas já no modelo, através de um teste t ou de um teste F e retirar aquelas que não sejam significativas. Retomar o passo 2.

Passo 6. Terminar com o modelo atual com v variáveis.

Note-se que o nível de significância utilizado nos testes para incluir novas variáveis deverá ser inferior ao nível de significância dos testes utilizados para retirar variáveis.

[9] Valor-P é a probabilidade de uma estatística de teste tenha, em relação ao valor observado, um valor extremo sob a hipótese nula H_0.

Capítulo 4

4. Modelo de Previsões de Consumos e os seus Resultados

4.1 Introdução

O modelo de previsões de consumos de energia elétrica em Portugal, tal como foi demonstrado no capítulo anterior, pode ser tratado como um modelo de regressão linear múltipla condicional. Deste modo, é hora de colocar em prática os resultados teóricos anteriormente apresentados.

O objetivo deste capítulo é descrever em detalhe os resultados obtidos na construção de vários modelos de regressão linear múltipla que melhor explicam a variável resposta pretendida. Ainda serão explicadas passo a passo as decisões que foram tomadas ao longo de processo de construção de um modelo de regressão linear múltipla, tal como as variáveis a serem incluídas. Por fim, serão explicados os modelos de RLM com melhores resultados.

4.2 Definição de Variáveis

Os consumos de energia elétrica dependem diariamente de múltiplas variáveis. Estas variáveis subdividem-se em duas categorias: variáveis externas e variáveis internas ao consumo diário.

As variáveis externas ao consumo são aquelas variáveis que dependem sobretudo das condições atmosféricas diariamente, tais como a temperatura, a precipitação, a humidade no ar, a velocidade do vento e, como, são sazonais, ou seja, têm um comportamento bem diferenciado no Verão e no Inverno.

As variáveis internas são definidas através dos consumos históricos. Estas variáveis ao serem caraterizadas pelo histórico, acabam por expressar algumas das variáveis externas, por exemplo expressam as variações de temperatura no dia em questão. Estas variáveis ainda têm em conta a sazonalidade do dia útil e fim-de-semana, tal como as quebras/aumentos de consumo em feriados/épocas festivas.

Para a construção das variáveis internas ao consumo e tendo em conta a natureza e características dos consumos diários, torna-se necessário assumir os seguintes pressupostos:

1. A disponibilização dos consumos históricos por parte da distribuidora de energia é de 1 dia;
2. As previsões de consumos de energia elétrica terão que ser realizadas dois dias antes do dia a prever, para se realizar a oferta de compra em mercado.
3. A existência de um comportamento bem diferenciado entre os consumos de dias úteis e fim-de-semana faz com que as previsões dos dias úteis só possam ter em conta o histórico de consumos (diários) de dias úteis e as previsões dos fins-de-semana só possam ter em conta histórico de consumos (diários) de dias "não" úteis.
4. A sazonalidade de feriados assemelha-se a um domingo, deste modo a previsão de um feriado tem em conta o domingo anterior mais próximo com consumos.
5. Não são considerados os consumos históricos dos feriados, em variáveis diárias.

Tendo em consideração os pressupostos 1 e 2 admite-se que para o dia a prever (D) os consumos históricos mais recentes são de há 3 dias atrás (D-3).

Assim, através destes pressupostos, definiram-se 8 variáveis internas ao consumo: Consumo D-3, Consumo D-4, Consumo D-5, Consumo D-6, Consumo D-7, Consumo Anual, Consumo Mensal e o Consumo Semanal.

O Consumo D-*j*, *j*= {3,4,5,6,7}, tem como referência o consumo diário de há *j* dias atrás. O Consumo Anual é definido pela seguinte equação:

$$Consumo\ Anual = \sum_{j=3}^{368} Consumo\ D-j \tag{4.1}$$

De modo semelhante, o consumo mensal e o consumo semanal são definidos através da equação 4.2. e equação 4.3, respetivamente.

$$Consumo\ Mensal\ = \sum_{j=3}^{33} Consumo\ D-j \tag{4.2}$$

$$Consumo\ Semanal\ = \sum_{j=3}^{9} Consumo\ D-j \tag{4.3}$$

As variáveis definidas como Consumo Anual, Consumo Mensal e Consumo Semanal não contêm nenhuma restrição em relação ao tipo de dia, útil ou fim-de-semana, nem em relação a feriados. Por outro lado, as variáveis Consumo D-*j*, *j*= {3,4,5,6,7} têm em consideração a restrições anteriormente mencionadas. Assim, para cada variável definiu-se o consumo histórico a usar para dia de semana.

Tabela 4.1: Dias de Referência para cada variável Consumo D-j, j= {3,4,5,6,7} – Descrição do dia de referência a ser utilizado para cada dia de semana para todas as variáveis Consumo D-j, j= {3,4,5,6,7}

Dia a prever	2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Sábado	Domingo
Consumo$_{D-3}$	6ª feira	6ª feira	6ª feira	2ª feira	3ª feira	Sábado	Domingo
Consumo$_{D-4}$	5ª feira	6ª feira	6ª feira	6ª feira	2ª feira	Sábado	Domingo
Consumo$_{D-5}$	4ª feira	5ª feira	6ª feira	2ª feira	3ª feira	Sábado	Domingo
Consumo$_{D-6}$	3ª feira	4ª feira	5ª feira	6ª feira	6ª feira	Sábado	Domingo
Consumo$_{D-7}$	2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Sábado	Domingo

A variável Consumo$_{D-3}$ tem como referência o consumo de há 3 dias atrás, para dias com a mesma tipologia útil/fim-de-semana. Deste modo, para prever uma terça-feira, a variável Consumo D-3, tinha como referência o consumo histórico de um Sábado. Nestes casos em que a tipologia da referência a ser usada é diferente da do dia a prever, utiliza-se como referência o histórico de consumos de dias com o mesmo tipo mais recente. Assim o dia de referência para uma terça-feira na variável Consumo$_{D-3}$ será uma sexta-feira, conforme se encontra descrito na tabela 4.1. É de salientar, que a referência de um sábado ou de um domingo a prever seja, respetivamente, um sábado ou um domingo.

Por último, quando o dia a prever é um feriado, qualquer que seja a variável Consumo D-j, j= {3,4,5,6,7} a referência é o consumo histórico do domingo mais recente. Quando a referência do consumo a ser utilizada seja um feriado, utiliza-se como referência o dia imediatamente anterior do mesmo género do dia a prever.

As variáveis externas ao consumo utilizadas foram a Temperatura Máxima, Temperatura Média, Temperatura Mínima e a Velocidade Média do Vento diária.

Definidas todas as variáveis a serem testadas na construção de um modelo de previsão de consumos de energia elétrica, é altura de descrever as unidades de cada uma das variáveis quer sejam internas ou externas ao consumo. Deste modo, as unidades para cada uma das variáveis encontra-se descrita na seguinte tabela 4.2.

Tabela 4.2: Unidade das variáveis internam e externas ao consumo de energia elétrica em Portugal

VARIÁVEIS	UNIDADES
CONSUMO D-J, J= {3,4,5,6,7}	GWh
CONSUMO ANUAL	GWh
CONSUMO MENSAL	GWh
CONSUMO SEMANAL	GWh
TEMPERATURA MÍNIMA	°C
TEMPERATURA MÉDIA	°C
TEMPERATURA MÁXIMA	°C
VELOCIDADE MÉDIA DO VENTO	km/h

4.3 Multicolinearidade

O estudo de possíveis problemas de multicolinearidade nos modelos para o consumo de energia elétrica será realizado através dos fatores de inflação das variâncias, da análise dos números condição e das componentes da variância, com o objetivo de identificar quais são as variáveis que tenham entre si uma relação demasiado forte e que, portanto, não devem ser utilizadas simultaneamente no mesmo modelo.

Considerem-se 2 modelos em estudo: o modelo com variáveis internas ao consumo e o modelo com variáveis internas e externas ao consumo. O modelo com variáveis internas tem em conta apenas as variáveis que dependem diretamente do consumo, tais como Consumo D-j, Consumo Anual, Consumo Mensal e Consumo Semanal. Por outro lado, o modelo com variáveis internas e externas ao consumo tem em conta para além das variáveis já referidas, as variáveis que não dependem do consumo: Temperatura Mínima, Temperatura Média, Temperatura Máxima e a Velocidade Média do Vento.

O estudo da multicolinearidade foi realizado tanto para o modelo que depende apenas de variáveis internas ao consumo, como para o modelo com variáveis internas e externas ao consumo.

4.3.1 Modelo com Variáveis Internas ao Consumo

Nesta secção serão apresentados os resultados obtidos no estudo da multicolinearidade no modelo com variáveis internas ao consumo de energia elétrica.

4.3.1.1 Fatores de Inflação da Variância

O estudo dos fatores de inflação da variância foi realizado através da construção da matriz inversa de correlações das variáveis independentes, uma vez que os elementos da diagonal desta matriz são exatamente os fatores de inflação em estudo.

A matriz inversa da matriz de correlações das variáveis independentes internas ao consumo encontra-se representada pela tabela 4.3[10].

Tabela 4.3: Primeira Iteração da Inversa da Matriz de Correlações para Variáveis Internas ao Consumo

Matriz Inversa de Correlações								
	Cons. D-3	Cons. D-4	Cons. D-5	Cons. D-6	Cons. D-7	Cons. Anual	Cons. Mensal	Cons. Semanal
Cons. D-3	19,79	-16,13	-2,16	0,42	-1,02	-0,01	1,11	-1,62
Cons. D-4	-16,13	39,91	-24,47	2,00	-0,58	0,11	0,17	-1,06
Cons. D-5	-2,16	-24,47	60,10	-35,31	1,53	-0,06	-0,71	1,03
Cons. D-6	0,42	2,00	-35,31	53,55	-20,75	-0,15	1,20	-0,82
Cons. D-7	-1,02	-0,58	1,53	-20,75	21,54	0,12	-1,87	1,26
Cons. Anual	-0,01	0,11	-0,06	-0,15	0,12	1,04	-0,25	0,07
Cons. Mensal	1,11	0,17	-0,71	1,20	-1,87	-0,25	3,73	-3,07
Cons. Semanal	-1,62	-1,06	1,03	-0,82	1,26	0,07	-3,07	4,43

O objetivo é obter um modelo em que não exista multicolinearidade entre as variáveis, ou seja, que a diagonal da matriz inversa da matriz de correlações seja não superior a 3. Assim, este processo é realizado até que na diagonal da matriz inversa de correlações apareçam valores não superiores a 3.

Deste modo, analisando a diagonal da matriz representada na tabela 4.3, a variável Consumo D-5 é a variável que contém o valor mais elevado na diagonal. Assim, a variável que representa do consumo de energia elétrica de há 5 dias atrás é retirada do modelo.

A última iteração em que a diagonal da matriz inversa das correlações apresenta valores não superiores a 3 está descrita na tabela 4.4.

Tabela 4.4: Última Iteração da Inversa da Matriz de Correlações para variáveis internas ao consumo

Matriz Inversa de Correlação			
	Cons. D-7	Cons. Anual	Cons. Mensal
Cons. D-7	1,53	0,02	-0,91
Cons. Anual	0,02	1,03	-0,20
Cons. Mensal	-0,91	-0,20	1,57

Por fim, conclui-se que as variáveis internas ao consumo de energia elétrica que não apresentam multicolinearidade são as variáveis Consumo D-7, Consumo Anual e Consumo Mensal, em que os valores da diagonal são todos inferiores a 3.

[10] CONS. - Consumo

4.3.1.2 Valores Próprios, Números Condição e Componentes da Variância

Os valores próprios e os respetivos números de condição tal como as componentes da variância para o modelo com variáveis internas ao consumo encontram-se descritos na tabela 4.5.

Tabela 4.5: Valores Próprios, Números de Condição e Proporções de Variância para variáveis internas ao consumo de energia elétrica em Portugal

Valores Próprios	Proporções de Variância									Números de Condição
	$B_{CONSTANTE}$	$B_{CONS.\ D-3}$	$B_{CONS.\ D-4}$	$B_{CONS.\ D-5}$	$B_{CONS.\ D-6}$	$B_{CONS.\ D-7}$	$B_{CONS.\ ANUAL}$	$B_{CONS.\ MENSAL}$	$B_{CONS.\ SEMANAL}$	
3,45E-04	0,00	0,43	0,05	0,13	0,04	0,43	0,00	0,01	0,01	161,37
8,98E+00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00
3,11E-05	1,00	0,00	0,00	0,00	0,00	0,00	1,00	0,01	0,00	537,05
1,00E-04	0,00	0,01	0,26	0,85	0,69	0,13	0,00	0,01	0,00	299,04
1,93E-04	0,00	0,35	0,65	0,02	0,24	0,24	0,00	0,00	0,00	215,87
9,51E-04	0,00	0,21	0,03	0,00	0,02	0,18	0,00	0,01	0,01	97,15
1,60E-02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	23,68
2,92E-03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,11	0,12	55,48
5,66E-04	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,85	0,85	125,92

É preciso ter em conta que valores próprios próximos de zero indicam a existência de dependência linear entre outras variáveis independentes. Tem-se ainda que um valor próprio ao qual corresponde um número condição superior a 30 indica também a existência de uma relação linear entre variáveis da matriz de planeamento.

Assim, conclui-se que para um valor próprio próximo de zero, com número condição superior ou igual a 30 e tal que as componentes da variância associadas estejam próximas da unidade, estamos na presença de uma relação aproximadamente linear entre algumas das variáveis em estudo.

Pela tabela 4.5 conclui-se a existência de uma relação entre as seguintes variáveis:

1. Constante e o Consumo Anual;
2. Consumo D-5 e o Consumo D-6;
3. Consumo D-3, Consumo D-4, Consumo D-6 e Consumo D-7;
4. Consumo Mensal e o Consumo Semanal.

A relação descrita em 1 pode ser explicada pelo facto do consumo anual sofrer apenas variações muito lentas, mantendo-se assim quase constante.

A relação entre o Consumo D-5 e o Consumo D-6 é explicada do mesmo modo que a relação entre as variáveis descritas em 1.

Por outro lado, a relação entre o Consumo Mensal e o Consumo Semanal deve-se ao facto de o Consumo Mensal ser construído com base em Consumos Semanais.

É necessário salientar mais uma vez que a base de construção de cada uma das variáveis é a mesma neste modelo de RLM, ou seja, o histórico de consumos de energia elétrica. Daí existir a relação descrita em 3.

Deste modo, é preciso ter atenção à inserção destas variáveis no modelo em estudo, uma vez que poderão provocar um mau ajustamento do modelo e provocar um aumento na variância dos estimadores dos coeficientes.

4.3.2 Modelo com Variáveis Internas e Variáveis Externas ao Consumo

Nesta secção serão demonstrados os resultados obtidos no estudo de multicolinearidade para o modelo com variáveis internas e externas ao consumo de energia elétrica.

4.3.2.1 Inflação da Variância

Tal como foi anteriormente referido, o estudo dos fatores de inflação da variância começa pela construção da inversa da matriz de correlações das variáveis independentes em estudo e tentar perceber quais são as variáveis com valores *VIF* superior a 3.

A primeira iteração da inversa da matriz de correlações das variáveis independentes internas e externas ao consumo encontra-se representada pela tabela 4.6.[11]

Tabela 4.6: Primeira Iteração da Inversa da Matriz de Correlações para Variáveis Internas e Externas ao Consumo

Inversa da matriz de correlações												
	Cons. D-3	Cons. D-4	Cons. D-5	Cons. D-6	Cons. D-7	Cons. Anual	Cons. Mensal	Cons. Semanal	Temp. Mínima	Temp. Média	Temp. Máxima	Vel. Med. Vento
Cons. D-3	19,43	-16,46	-1,97	0,19	-0,57	-0,02	0,64	-0,62	-0,20	-0,12	0,61	-0,10
Cons. D-4	-16,46	39,88	-24,42	1,99	-0,30	0,12	0,20	-1,01	-0,18	1,00	-0,86	-0,01
Cons. D-5	-1,97	-24,42	60,40	-35,39	1,17	-0,10	-1,13	0,83	0,41	-2,40	1,50	0,14
Cons. D-6	0,19	1,99	-35,39	53,44	-20,33	-0,10	0,92	-0,35	0,77	-0,45	-0,02	-0,03
Cons. D-7	-0,57	-0,30	1,17	-20,33	20,97	0,11	-0,48	-0,28	-0,96	2,28	-1,40	-0,05
Cons. Anual	-0,02	0,12	-0,10	-0,10	0,11	1,06	-0,24	0,06	0,35	-0,27	-0,08	0,04
Cons. Mensal	0,64	0,20	-1,13	0,92	-0,48	-0,24	4,03	-2,95	1,08	-0,41	0,29	-0,14
Cons. Semanal	-0,62	-1,01	0,83	-0,35	-0,28	0,06	-2,95	4,55	-0,20	0,36	0,04	0,06
Temp. Mínima	-0,20	-0,18	0,41	0,77	-0,96	0,35	1,08	-0,20	22,01	-38,27	19,27	-0,80
Temp. Média	-0,12	1,00	-2,40	-0,45	2,28	-0,27	-0,41	0,36	-38,27	79,18	-44,63	0,64
Temp. Máxima	0,61	-0,86	1,50	-0,02	-1,40	-0,08	0,29	0,04	19,27	-44,63	28,15	-0,08
Vel. Med. Vento	-0,10	-0,01	0,14	-0,03	-0,05	0,04	-0,14	0,06	-0,80	0,64	-0,08	1,08

Tal como foi dito anteriormente, o objetivo é obter um modelo em que não haja multicolinearidade entre as variáveis em estudo. A análise da diagonal da matriz inversa de correlações indica quais as variáveis que contêm uma relação entre as restantes variáveis do modelo.

Analisando a tabela 4.6, a variável com valor mais elevado na diagonal da matriz inversa de correlações é a Temperatura Média Diária. Deste modo dever-se-á excluir esta variável no modelo de regressão múltipla, uma vez que esta pode ser escrita aproximadamente como uma função linear de algumas das restantes variáveis.

O processo é realizado sucessivamente até que a diagonal contenha apenas valores não superiores a 3. A última iteração da análise de multicolinearidade às variáveis internas e externas ao consumo de energia elétrica é composta pelas variáveis Consumo D-7, Consumo Anual, Consumo Mensal, Temperatura Máxima e Velocidade Média do Vento, como demonstra a tabela 4.7, não existindo assim relação entre estas variáveis.

[11] CONS. – Consumo, TEMP_MIN – Temperatura Mínima, TEMP_MED – Temperatura Média, TEMP_MAX – Temperatura Máxima, VEL_VENTO – Velocidade Média do Vento

Tabela 4.7: Última Iteração da Inversa da Matriz de Correlações para variáveis internas e externas ao consumo

Inversa da Matriz de Correlação					
	Consumo D-7	Consumo Anual	Consumo Mensal	Temperatura Máxima	Velocidade Média Vento
Consumo D-7	1,43	0,03	-0,75	0,06	-0,03
Consumo Anual	0,03	1,05	-0,25	-0,09	0,07
Consumo Mensal	-0,75	-0,25	1,89	0,78	0,01
Temperatura Máxima	0,06	-0,09	0,78	1,45	-0,04
Velocidade Média Vento	-0,03	0,07	0,01	-0,04	1,01

4.3.2.2 Valores Próprios, Números Condição e Componentes da Variância

Nesta secção serão apresentados os resultados obtidos na análise de multicolinearidade às variáveis internas e externas ao consumo de energia elétrica através do estudo dos valores próprios, números condição e componentes da variância.

Como já foi referido anteriormente, valores próprios próximos de zero, com número condição não inferior a 30 e com componentes da variância próximas da unidade indicam a existência de uma relação linear entre as variáveis em estudo.

Os valores próprios, os números condição e as componentes da variância para as variáveis internas e externas ao consumo encontram-se descritos na tabela 4.8.

Tabela 4.8: Valores Próprios, Números de Condição e Proporções da Variância para as variáveis internas e externas ao consumo de energia elétrica em Portugal

Valores Próprios	Proporções de Variância													Números de Condição
	$B_{CONSTANTE}$	$B_{CONS.\ D\text{-}3}$	$B_{CONS.\ D\text{-}4}$	$B_{CONS.\ D\text{-}5}$	$B_{CONS.\ D\text{-}6}$	$B_{CONS.\ D\text{-}7}$	$B_{CONS.\ ANUAL}$	$B_{CONS.\ MENSAL}$	$B_{CONS.\ SEMANAL}$	B_{TEMP_MIN}	B_{TEMP_MED}	B_{TEMP_MAX}	B_{VEL_VENTO}	
3,42E-04	0,00	0,43	0,05	0,13	0,04	0,42	0,00	0,01	0,01	0,01	0,01	0,01	0,00	192,03
1,26E+01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00
1,92E-04	0,00	0,35	0,65	0,02	0,24	0,25	0,00	0,00	0,00	0,00	0,00	0,00	0,00	255,96
3,03E-05	0,99	0,00	0,00	0,00	0,00	0,00	0,99	0,00	0,00	0,01	0,00	0,00	0,00	645,22
1,00E-04	0,00	0,01	0,26	0,85	0,69	0,13	0,00	0,01	0,00	0,00	0,00	0,00	0,00	354,80
9,46E-04	0,00	0,21	0,03	0,00	0,02	0,19	0,00	0,01	0,01	0,01	0,01	0,00	0,00	115,38
1,30E-02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,02	31,12
1,75E-03	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,09	0,21	0,01	0,00	0,00	0,00	84,86
5,42E-04	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,85	0,74	0,05	0,03	0,02	0,00	152,39
2,44E-01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	7,18
6,93E-04	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,02	0,77	0,95	0,88	0,00	134,84
1,83E-02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,14	0,00	0,07	0,06	26,25

Analisando a tabela 4.8, é evidente uma relação entre as seguintes variáveis:

1. Consumo D-3 e o Consumo D-7;
2. Consumo D-3, Consumo D-4, Consumo D-6 e Consumo D-7;
3. Constante e o Consumo Anual;
4. Consumo D-4, Consumo D-5, Consumo D-6 e Consumo D-7;
5. Consumo Mensal e Consumo Semanal;
6. Temperatura Mínima, Temperatura Média e Temperatura Máxima.

As relações descritas em 1 até 5 justificam-se de modo semelhante ao que foi feito em 4.3.1.2.

A relação entre as variáveis Temperatura Mínima, Média e Máxima deve-se ao facto de a temperatura média ser construída com base nas restantes variáveis.

Deste modo, é preciso ter atenção à inserção destas variáveis no modelo em estudo, uma vez que poderão provocar um mau ajustamento do modelo e um aumento na variância dos estimadores dos coeficientes.

É de esperar que as relações obtidas apenas com variáveis internas ao consumo se mantenham em ambos os modelos, como se pode comprovar pela tabela 4.8 onde se verifica a existência de uma relação entre as variáveis descritas em 2, 3 e 5.

4.4 Regressão Linear Múltipla

Através dos resultados obtidos no capítulo 3 conclui-se que uma vez verificadas as condições do teorema de máxima verosimilhança, o modelo de previsão de consumos de energia elétrica em Portugal pode ser tratado como um modelo de regressão múltipla. Este modelo de previsão tem a particularidade de as variáveis dependentes serem também variáveis independentes do modelo de regressão.

Tal como já foi referido, os estimadores obtidos pelo método da máxima verosimilhança são exatamente os mesmos que são obtidos pelo método dos mínimos quadrados. Uma vez que é usual estimar os coeficientes das variáveis independentes em modelos de regressão múltipla pelo método dos mínimos quadrados, este método será utilizado para estimar os parâmetros b_1, b_2, …, b_m, isto é pelos valores que produzem o subespaço linear gerado pelas variáveis independentes tal que a soma dos quadrados das distâncias das observações a esse espaço seja mínima. Assim, tal como foi referido no capítulo 3, os coeficientes de cada uma das variáveis, tal como a constante são gerados através dos Estimadores de Mínimos Quadrados.

Deste modo, o objetivo desta secção é a criação de modelos de regressão linear múltipla com o fim de prever consumos de energia elétrica em Portugal Continental.

A construção dos modelos é feita tendo em conta as variáveis obtidas através do método de seleção de variáveis *Stepwise* e Regressivo. Este estudo irá ser repartido em duas partes. Na primeira, a criação de um modelo de previsão de consumos depende apenas de variáveis internas ao consumo e, na segunda, ter-se-á em conta, para além de variáveis internas, variáveis externas ao consumo. Para qualquer uma das partes ir-se-á realizar a criação dos modelos através de ambos os métodos de seleção de variáveis.

Antes de se iniciar a construção dos modelos de RLM, é necessário tentar perceber o comportamento de cada uma das variáveis em estudo. Deste modo, foi realizada uma análise às variáveis em estudo para o ano de 2015.

A evolução do consumo diário para as variáveis Consumo D-3, Consumo D-4, Consumo D-5, Consumo D-6 e Consumo D-7 em Portugal ao longo de 2015 encontra-se descrito no gráfico 4.1[12]. Tal como é previsto, não existe uma discrepância muito acentuada entre estas variáveis uma vez que estas estão apenas desfasadas alguns dias segundo os pressupostos definidos na tabela 4.1.

Em Portugal, a variação do consumo anual em 2015 encontra-se ilustrada no gráfico 4.2[13]. O decréscimo do consumo de energia elétrica entre abril e maio deve-se à pouca necessidade de consumo de energia elétrica em aparelhos de aquecimento/arrefecimento de temperatura ambiente e ao aumento de horas diárias com luz solar. Por outro lado, a partir de maio o consumo de energia elétrica sofre um aumento até setembro devido à utilização de equipamentos de arrefecimento da temperatura ambiente.

[12] Fonte: Estatística Diária – SEN, REN
[13] Fonte: Estatística Diária – SEN, REN

A variação diária do Consumo Semanal e Mensal é pouco significativa tal como se pode comprovar pelo gráfico 4.3[14].

Ao longo do ano é de se esperar que as temperaturas diárias sejam mais altas na época de Verão e mais baixas no Inverno, e o ano 2015 não foi um ano que fugiu à regra, como se pode visualizar no gráfico 4.4[15]. Com isto é também presumível que a velocidade do vento tenha picos mais elevados em épocas com temperaturas baixas. Pelo gráfico 4.5[16], é evidente o aumento de velocidade média do vento entre janeiro e maio de 2015.

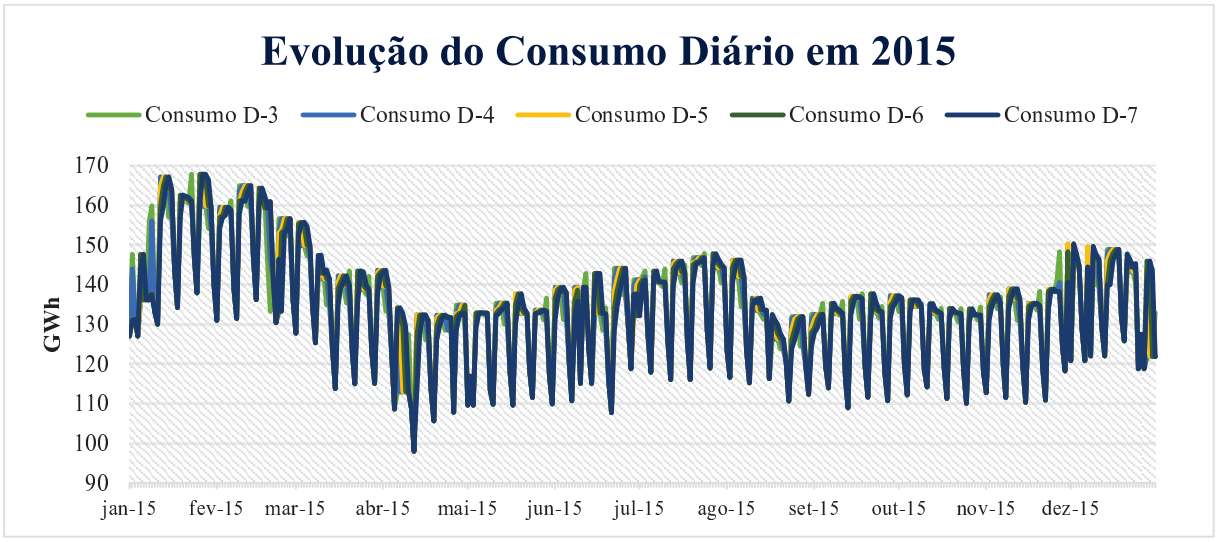

Gráfico 4.1: Evolução do consumo diário em 2015 em Portugal das variáveis Consumo D-j, j={3,4,5,6,7}

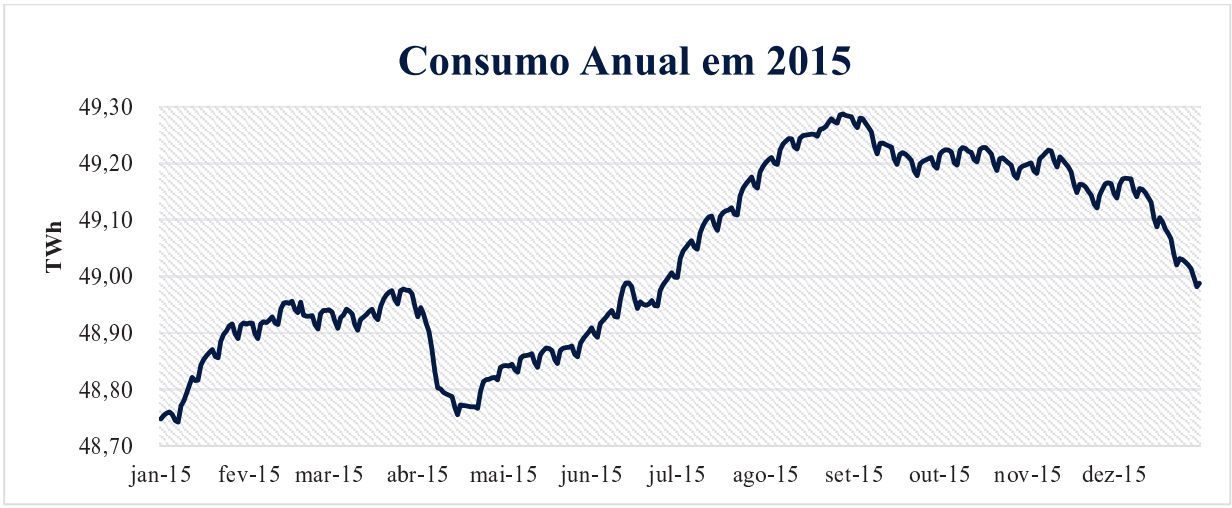

Gráfico 4.2: Evolução da variável Consumo Anual em Portugal durante 2015

[14] Fonte: Estatística Diária – SEN, REN
[15] Fonte: Resumo Mensal, Tempo em Lisboa
[16] Fonte: Resumo Mensal, Tempo em Lisboa

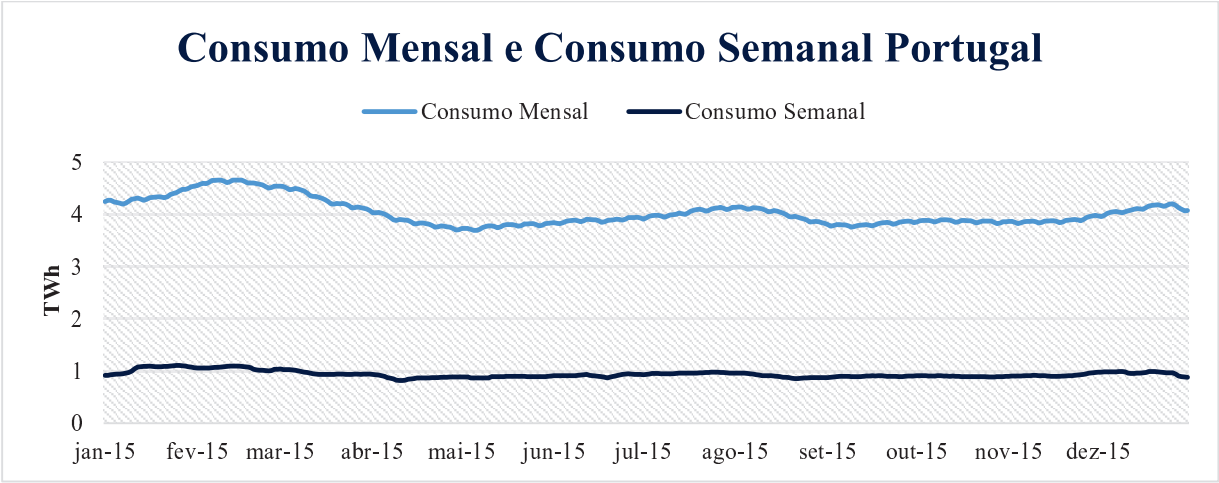

Gráfico 4.3: Evolução do Consumo Mensal e Consumo Semanal em Portugal durante o ano 2015

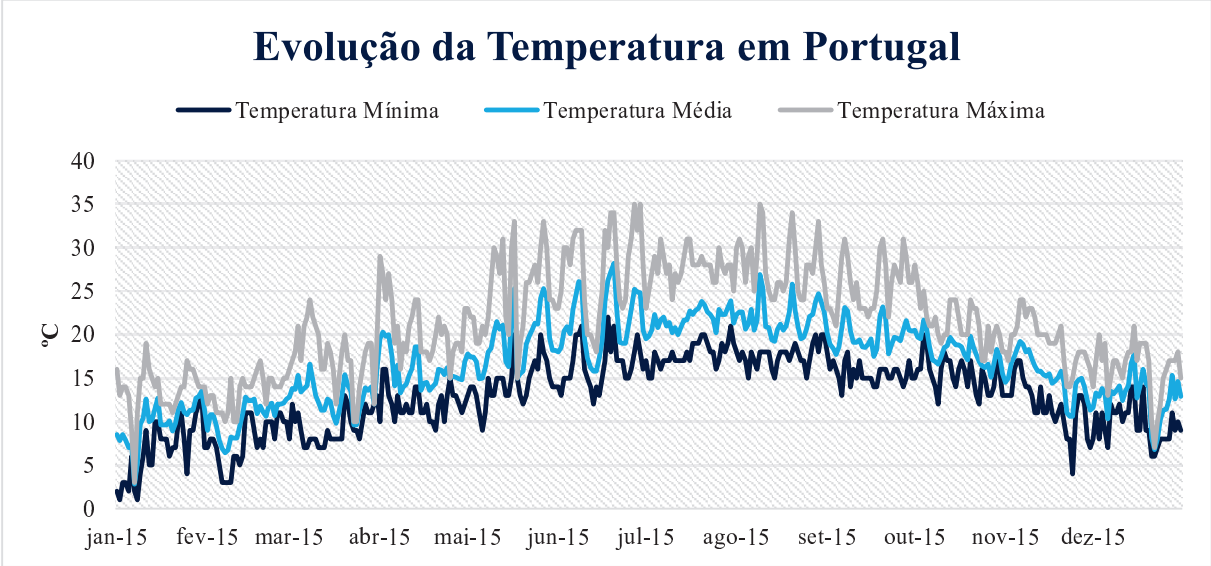

Gráfico 4.4: Evolução da Temperatura Mínima, Média e Máxima em Portugal ao longo de 2015

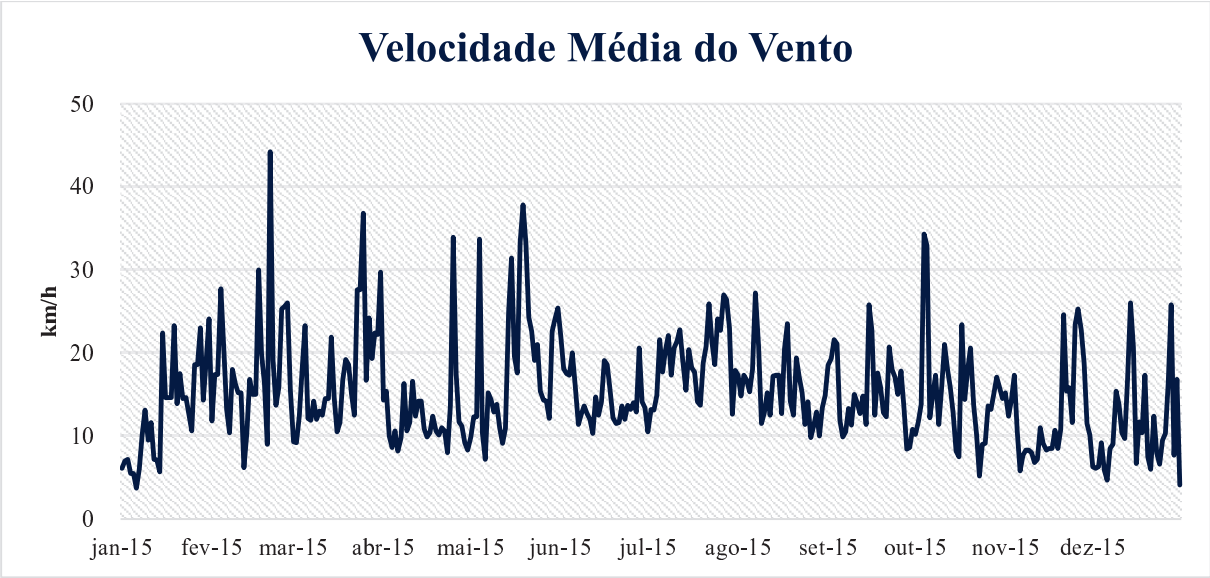

Gráfico 4.5: Evolução da Velocidade Média do Vento em Portugal em 2015

4.4.1 Modelo Com Variáveis Internas ao Consumo

As variáveis usadas nos métodos de seleção de variáveis *Stepwise* e Regressivo nesta secção são as variáveis internas ao consumo: Consumo D-j, j= {3,4,5,6,7}, Consumo Anual, Consumo Mensal e Consumo Semanal.

Em ambos os métodos de seleção de variáveis, o modelo de regressão linear múltiplo é o mesmo, tendo em conta apenas variáveis internas ao consumo. Este modelo encontra-se descrito na tabela 4.9, tal como os coeficientes obtidos através dos EMQ e as suas principais caraterísticas.

Tabela 4.9: Resultados do ajustamento de um modelo de regressão linear múltipla à previsão de consumos de energia elétrica como função de variáveis internas ao consumo obtido através do método de seleção de variáveis *Stepwise* e Regressivo

	$B_{CONSTANTE}$	$B_{CONSUMO\ D-3}$	$B_{CONSUMO\ D-7\ CONSUMO\ MENSAL}$		
Coefs.	20,09	0,57	0,41	-4,04E-03	Coefs.
Desv.Pad.	2,44	0,03	0,03	7,20E-04	Desv.Pad.
R^2	0,83	5,55	#N/A	#N/A	S
Estatistica F	2407,70	1457,00	#N/A	#N/A	G.L.
SQreg	222237,40	44828,46	#N/A	#N/A	SQe

As variáveis que melhor descrevem o modelo de regressão linear múltipla são o Consumo D-3, Consumo D-7 e Consumo Mensal. Estas variáveis fazem com que o modelo de regressão tenha um ajustamento consideravelmente bom, uma vez que o coeficiente de determinação R^2 é 0,83 para uma variância de aproximadamente 30,77.

O modelo matemático de regressão linear múltipla que descreve diariamente a previsão de consumos de energia elétrica em Portugal Continental é:

$$y = 20{,}09 + 0{,}57x_{Consumo\ D-3} + 0{,}41x_{Consumo\ D-7} - 0{,}00404x_{Consumo\ Mensal} \quad (4.4)$$

4.4.2 Modelo Com Variáveis Internas e Externas ao Consumo

As variáveis utilizadas na criação de um modelo de regressão linear múltipla nesta secção são as variáveis definidas como sendo internas e externas ao consumo de energia elétrica, isto é, Consumo D-j, j= {3,4,5,6,7}, Consumo Anual, Consumo Mensal, Consumo Semanal, Temperatura Mínima, Temperatura Média, Temperatura Máxima e Velocidade Média do Vento.

O modelo de regressão linear múltiplo obtido através do método de seleção *Stepwise* tendo em conta variáveis internas e externas ao consumo de energia elétrica encontra-se descrito na tabela 4.10, tal como os coeficientes obtidos através dos EMQ e as suas principais caraterísticas.

Tabela 4.10: Resultado do ajustamento de regressão linear múltipla à previsão de consumo de energia elétrica como função de variáveis internam e externas ao consumo obtido através do método de seleção de variáveis *Stepwise*

	$B_{CONSTANTE}$	$B_{CONSUMO\ D-3}$	$B_{CONSUMO\ D-7\ CONSUMO\ MENSAL}$	B_{TEMP_MIN}	
Coefs.	34,17	0,57	0,41	-0,01	-0,23	Coefs.
Desv.Pad.	3,43	0,03	0,03	0,00	0,04	Desv.Pad.
R^2	0,84	5,49	#N/A	#N/A	#N/A	S
Estatistica F	1854,10	1456,00	#N/A	#N/A	#N/A	G.L.
SQreg	223239,10	43826,76	#N/A	#N/A	#N/A	SQe

As variáveis que melhor descrevem o modelo de regressão linear múltipla com variáveis internas e externas ao consumo são o Consumo D-3, Consumo D-7 e Consumo Mensal e a Temperatura Mínima. Estas variáveis fazem com que o modelo de regressão tenha um ajustamento bom, uma vez que o coeficiente de determinação R^2 é 0,84 para uma variância de aproximadamente 5,49.

O modelo matemático de regressão linear múltipla que descreve diariamente a previsão de consumos de energia elétrica em Portugal Continental, tendo em conta variáveis internas e externas ao consumo através do método de seleção de variáveis *Stepwise*, é:

$$y = 34{,}17 + 0{,}57x_{Consumo\ D-3} + 0{,}41x_{Consumo\ D-7} - 0{,}01x_{Consumo\ Mensal} - 0{,}23x_{Temperatura\ Mínima} \quad (4.5)$$

Por outro lado, o modelo de RLM obtido através do método de seleção de variáveis Regressivo encontra-se descrito na tabela 4.11, tal como os coeficientes obtidos através dos EMQ e as suas principais caraterísticas.

Tabela 4.11: Resultado do ajustamento de regressão linear múltipla à previsão de consumo de energia elétrica como função de variáveis internam e externas ao consumo obtido através do método de seleção de variáveis Regressivo

	$B_{CONSTANTE}$	$B_{CONSUMO\ D-3}$	$B_{CONSUMO\ D-7\ CONSUMO\ MENSAL}$	B_{TEMP_MIN}	B_{TEMP_MED}	B_{TEMP_MAX}	
Coefs.	34,04	0,56	0,41	-0,01	-0,57	0,69	-0,33	Coefs.
Desv.Pad.	3,50	0,03	0,03	0,00	0,14	0,26	0,12	Desv.Pad.
R^2	0,84	5,48	#N/A	#N/A	#N/A	#N/A	#N/A	S
Estatistica F	1241,73	1454,00	#N/A	#N/A	#N/A	#N/A	#N/A	G.L.
SQreg	223456,42	43609,44	#N/A	#N/A	#N/A	#N/A	#N/A	SQe

As variáveis que melhor descrevem o modelo de RLM através do método de seleção Regressivo são o Consumo D-3, Consumo D-7 e Consumo Mensal, Temperatura Mínima, Temperatura Média e Temperatura Máxima. Estas variáveis fazem com que o modelo de regressão tenha um ajustamento

consideravelmente bom, uma vez que o coeficiente de determinação R^2 é 0,84 para uma variância de aproximadamente 5,48.

O modelo matemático de regressão linear múltipla que descreve diariamente a previsão de consumos de energia elétrica em Portugal Continental, tendo em conta variáveis internas e externas ao consumo através do método de seleção de variáveis Regressivo é:

$$y = 34{,}04 + 0{,}56x_{Consumo\ D-3} + 0{,}41x_{Consumo\ D-7} - 0{,}01x_{Consumo\ Mensal} - 0{,}57x_{Temperatura\ Mínima} + 0{,}69x_{Temperatura\ Média} - 0{,}33x_{Temperatura\ Máxima} \quad (4.6)$$

Neste último modelo de regressão linear múltipla existem variáveis com multicolinearidade entre si. Deste modo, não é aconselhável o uso deste modelo para prever consumos de energia elétrica em Portugal.

Assim, comparando o modelo de regressão linear múltipla com variáveis internas ao consumo e o modelo de RLM descrito na equação 4.5, este último modelo apresenta um melhor ajustamento aos valores observados.

As análises realizadas nas próximas secções têm em conta os três modelos, apesar de se ter chegado à conclusão que um deles não é aconselhável usar para prever consumos de energia elétrica em Portugal.

Assim, para simplificar, designa-se a partir de agora como modelo A o modelo de regressão linear múltipla apenas com variáveis internas ao consumo descrito na equação 4.4. O modelo B é definido pelo modelo de RLM em que o consumo de energia elétrica em Portugal é descrito através da ordenada na origem (Constante), Consumo D-3, Consumo D-7, Consumo Mensal e a Temperatura Mínima Diária em Portugal definido na equação 4.5. Por último, o modelo C é o modelo de regressão múltipla constituído pela Constante e pelas variáveis independentes Consumo D-3, Consumo D-7, Consumo Mensal, Temperatura Mínima e Máxima diária em Portugal descrito na equação 4.6. As caraterísticas dos 3 modelos anteriormente definidos encontram-se descritos na tabela 4.12.

Tabela 4.12: Caraterísticas dos modelos de regressão linear múltipla Finais

Caraterísticas dos Modelos de Regressão Múltipla Finais			
Modelos	Variáveis	R^2	S
A	Constante Consumo D-3 Consumo D-7 Consumo Mensal	0,83	5,55
B	Constante Consumo D-3 Consumo D-7 Consumo Mensal Temperatura Mínima Diária	0,84	5,49
C	Constante Consumo D-3 Consumo D-7 Consumo Mensal Temperatura Mínima Diária Temperatura Média Diária Temperatura Máxima Diária	0,84	5,48

4.5 Análise dos Resíduos

Os resíduos são definidos pela diferença entre os valores observados e os valores ajustados e descrevem as disparidades entre a realidade e o modelo teórico construído.

Tal como já foi descrito anteriormente, a análise dos resíduos é constituída pelas seguintes etapas:

1. Representação gráfica dos resíduos contra cada uma das variáveis independentes incluídas no modelo de regressão linear múltiplo;
2. Representação gráfica dos resíduos contra outras variáveis independentes que não tenham sido incluídas no modelo;
3. Representação gráfica dos resíduos contra os valores ajustados;
4. Estudo da normalidade dos resíduos através de histogramas e testes de ajustamento.

Nesta secção será demonstrada cada uma das etapas definidas a cima para cada um dos modelos obtidos A, B e C.

4.5.1 Análise dos Resíduos do Modelo A

O modelo A é constituído pelas variáveis Consumo D-3, Consumo D-7 e Consumo Mensal. A representação gráfica dos resíduos contra cada uma destas variáveis não demonstra a existência de uma relação entre elas. Conclui-se assim que não é necessário fazer nenhuma transformação às variáveis inseridas no modelo, como se pode visualizar nos gráficos 4.6, 4.7 e 4.8.

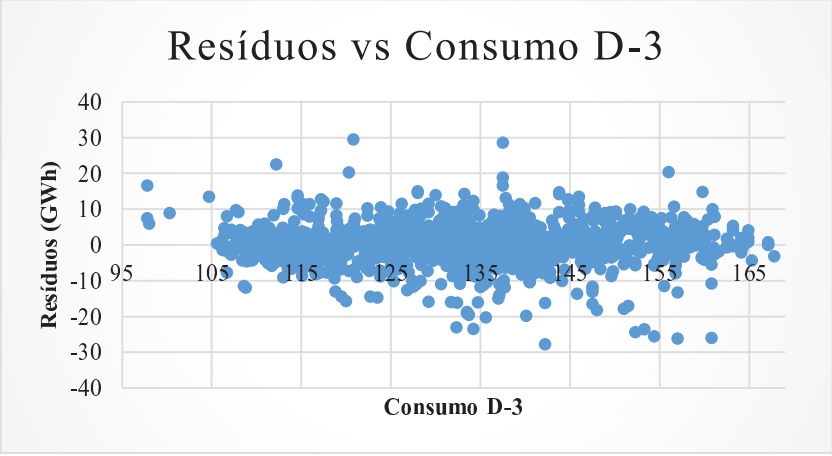

Gráfico 4.6: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo de há 3 dias atrás

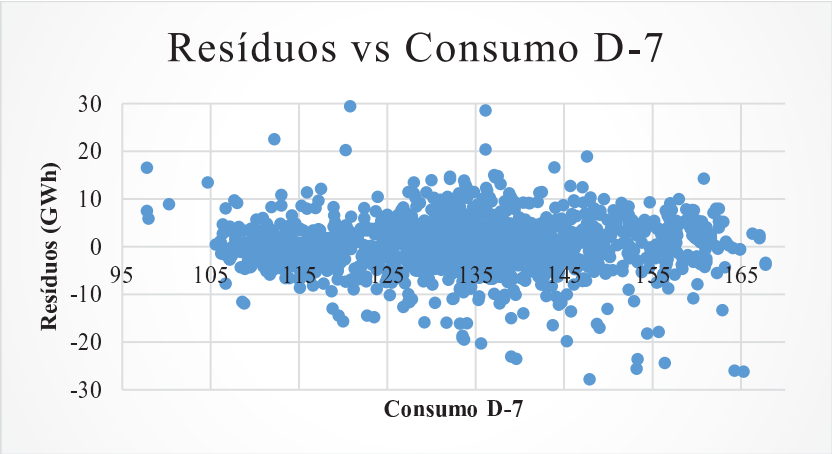

Gráfico 4.7: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo D-7

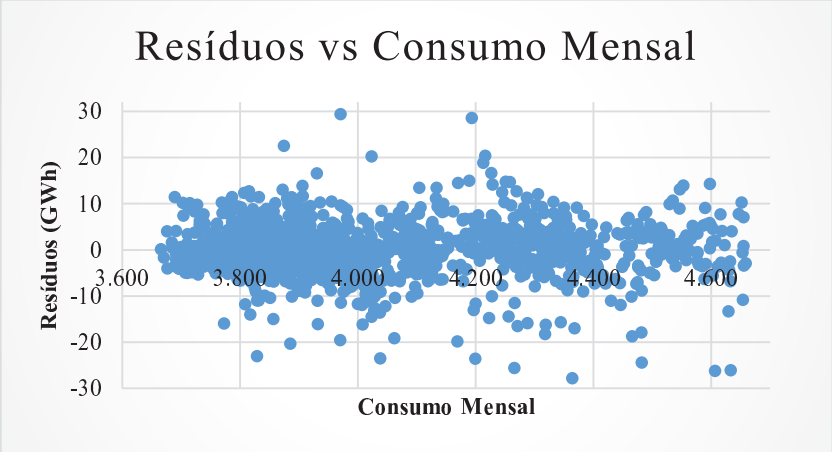

Gráfico 4.8: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo Mensal

A representação gráfica dos resíduos contra as restantes variáveis independentes que não foram consideradas no modelo de regressão linear múltipla A, Consumo D-4, Consumo D-5, Consumo D-6, Consumo Anual e Consumo Semanal encontra-se representada nos gráficos 4.9 a 4.13.

Para cada uma das variáveis independentes que não foi incluída no modelo A, a análise gráfica dos resíduos reforça a ideia de estas variáveis não terem impacto na previsão de consumos de energia elétrica em Portugal, uma vez que não existe uma relação entre cada uma destas variáveis e os resíduos obtidos do modelo A.

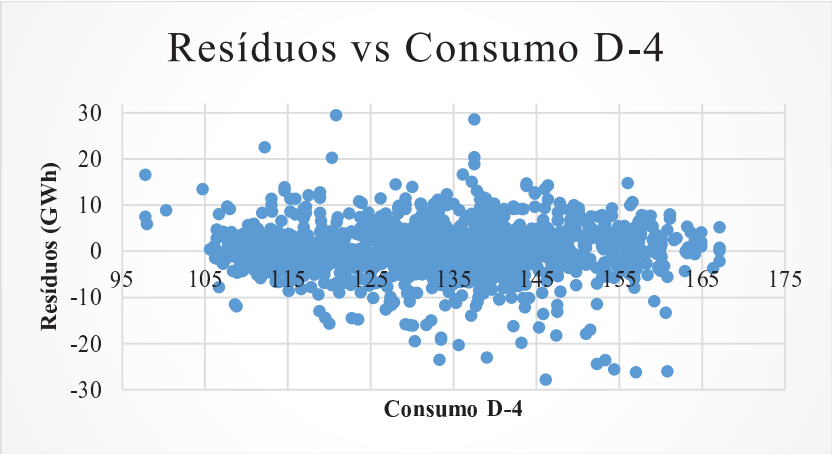

Gráfico 4.9: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo D-4

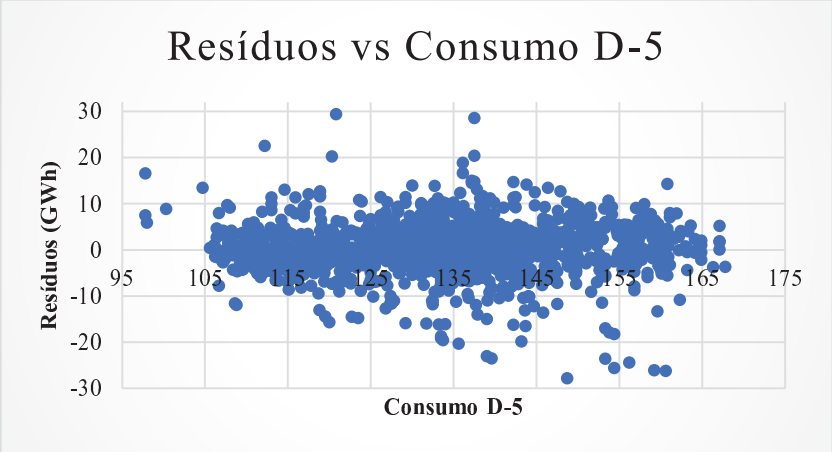

Gráfico 4.10: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo de há 5 dias atrás

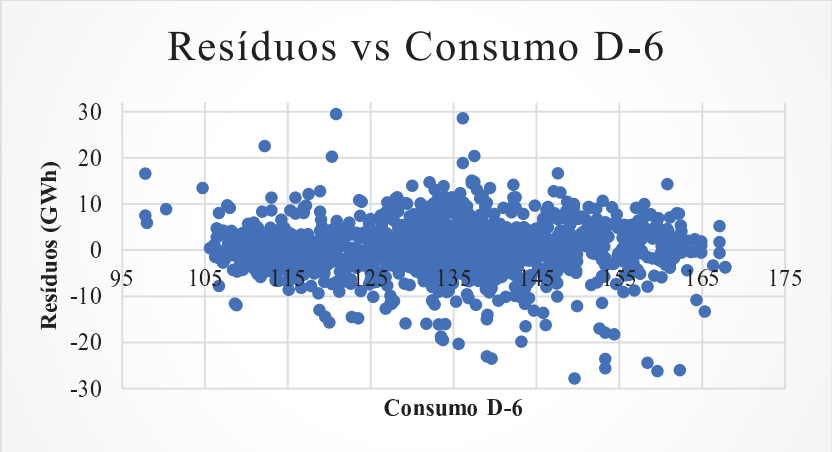

Gráfico 4.11: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo D-6

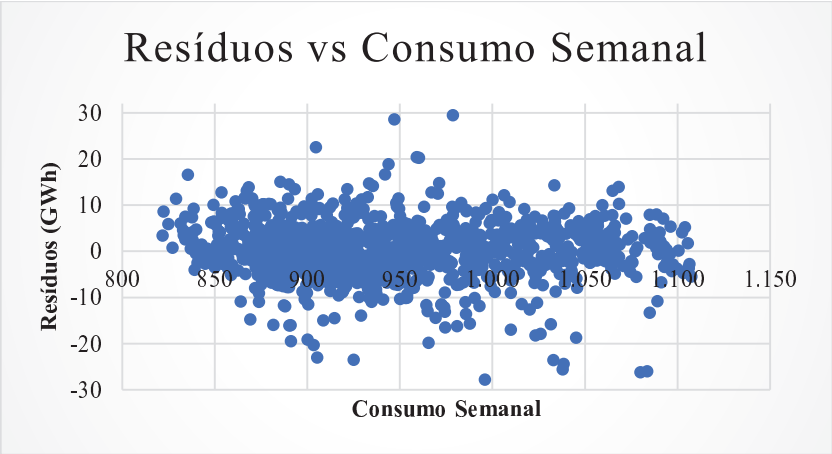

Gráfico 4.12: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo Semanal

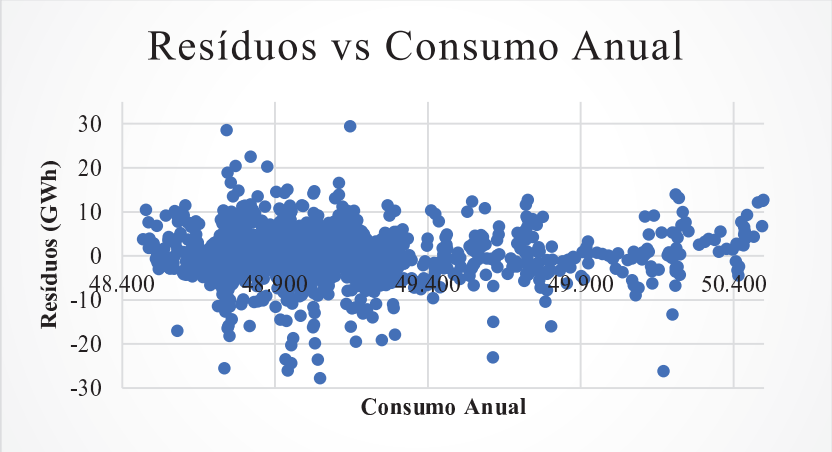

Gráfico 4.13: Representação Gráfica dos Resíduos do modelo A contra a variável Consumo Anual

O terceiro passo da análise dos resíduos é a representação gráfica destes contra os valores ajustados do modelo A. As variáveis inseridas no modelo de RLM A não necessitam de nenhuma transformação como se pode confirmar pelo gráfico 4.14.

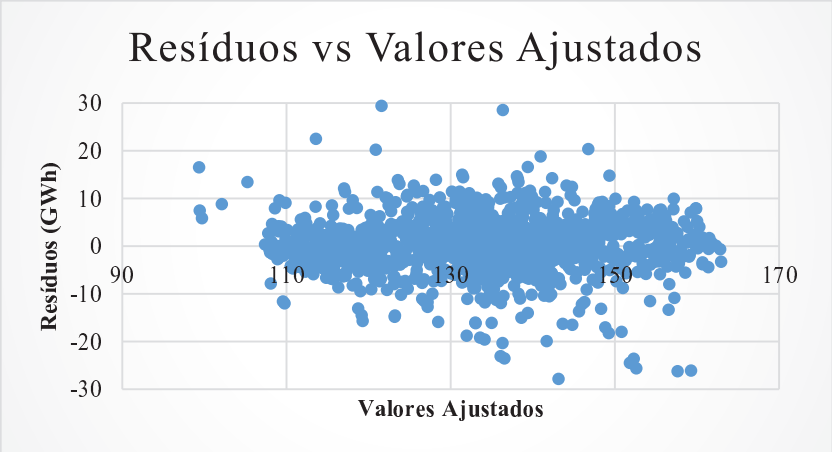

Gráfico 4.14: Representação Gráfica dos Resíduos contra os Valores Ajustados do modelo A

Por último, o estudo da normalidade dos resíduos foi realizado com o Teste de *Lilliefors*. Assim, as hipóteses de teste são:

H_0: Os resíduos seguem uma distribuição normal *versus* H_1:Os resíduos não seguem uma distribuição normal

Os resíduos do modelo de regressão linear múltipla A são definidos pelas seguintes caraterísticas amostrais, definidas na tabela 4.13:

Tabela 4.13: Principais caraterísticas amostrais do modelo de regressão múltipla A

Caraterísticas amostrais do Modelo A					
Tamanho da Amostra, *n*	Média	Desvio Padrão	Nível de Significância, α	d	$dc_{n,\propto}$[17]
1461	0	12,17	0,05	0,99	0,02

Como $d > dc_{1461,0.05}$, a hipótese nula é rejeitada, ou seja, rejeita-se a hipótese de os resíduos seguirem uma distribuição normal.

Esta conclusão é reforçada com a construção do histograma para os resíduos do modelo A, como se poderá visualizar no gráfico 4.15, em que as frequências relativas dos resíduos não seguem uma distribuição normal.

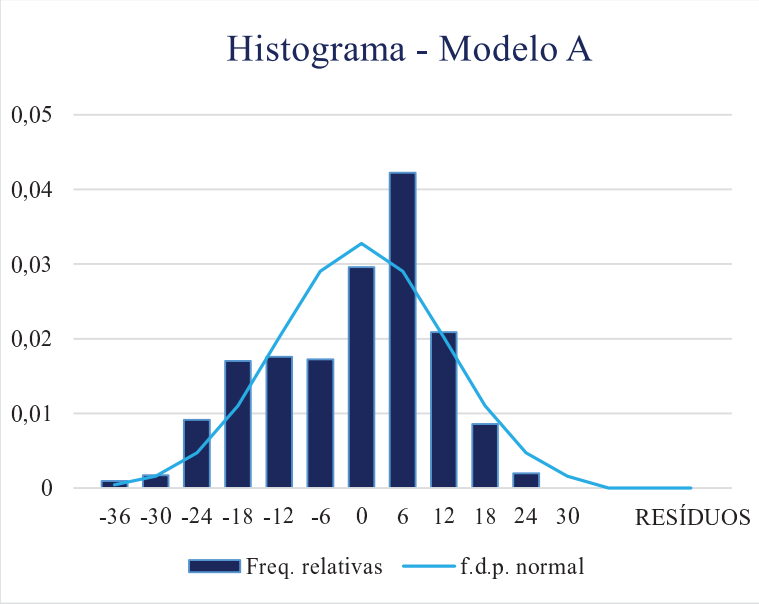

Gráfico 4.15: Representação gráfica dos Resíduos do modelo A e da função de distribuição da Normal

4.5.2 Análise dos Resíduos do Modelo B

O modelo de regressão linear múltipla B é constituído pelas variáveis Consumo D-3, Consumo D-7, Consumo Mensal e Temperatura Mínima diária.

[17] Ver Anexo 1 -Tabela de Estatísticas de Teste de *Lilliefors* para a distribuição Normal

As representações gráficas 4.16 a 4.19 negam uma relação entre os resíduos e cada uma das variáveis independentes incluídas no modelo de RLM B. Deste modo, não é necessário transformar nenhuma das variáveis independentes.

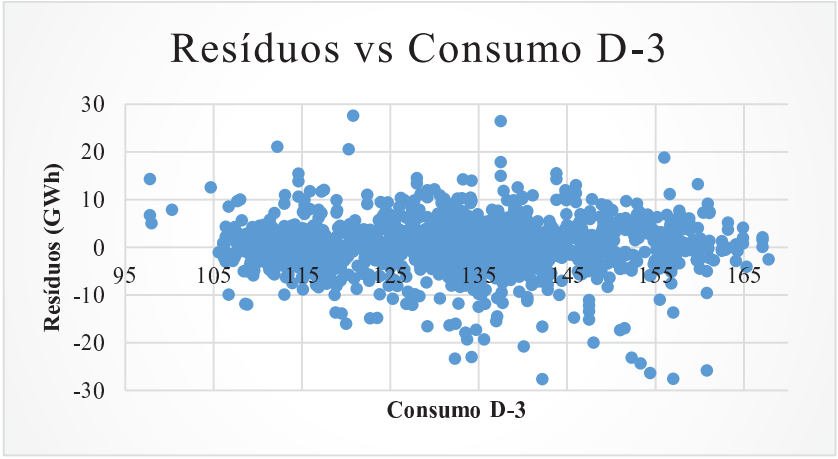

Gráfico 4.16: Representação gráfica dos Resíduos do modelo B *versus* Consumo de Energia Elétrica de há 3 dias atrás

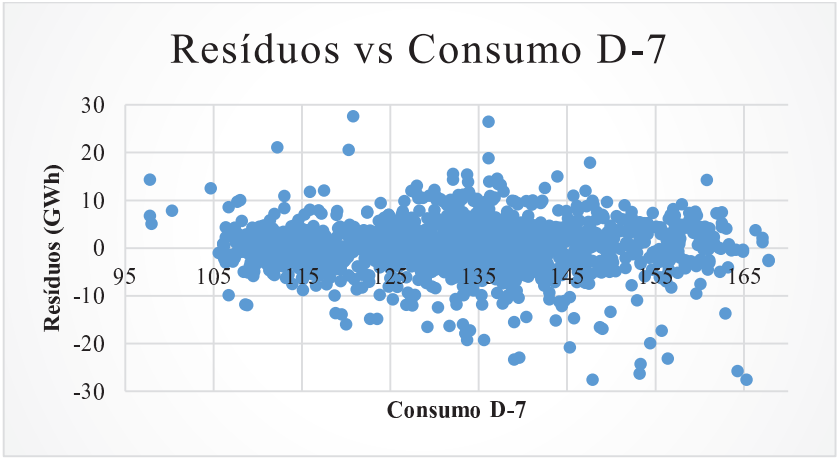

Gráfico 4.17: Representação gráfica dos Resíduos do modelo B contra a variável Consumo D-7

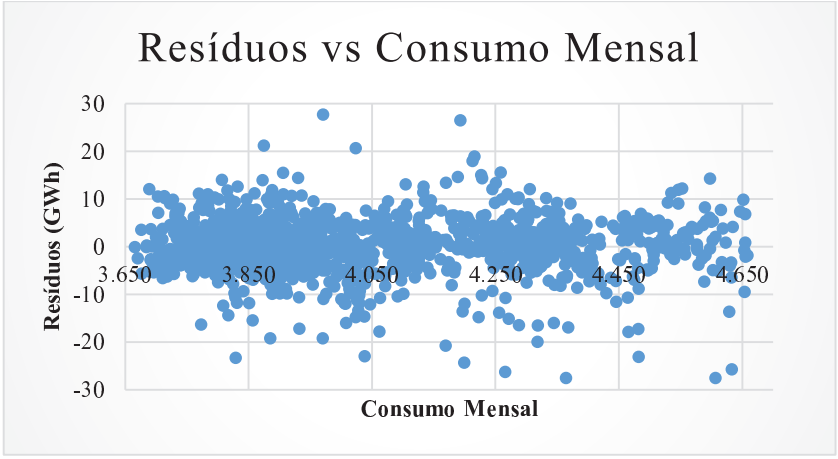

Gráfico 4.18: Representação gráfica dos Resíduos do modelo B contra a variável Consumo Mensal

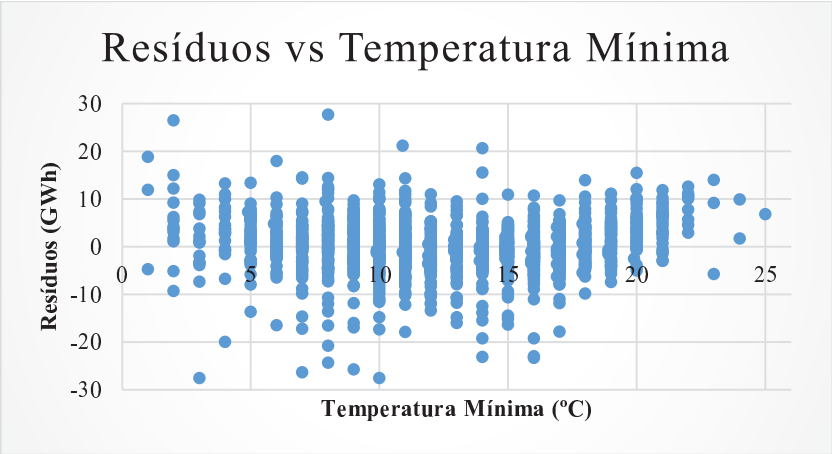

Gráfico 4.19: Representação gráfica dos Resíduos do modelo B contra a variável Temperatura Mínima em Portugal (ºC)

A representação gráfica dos resíduos contra as variáveis que não foram incluídas no modelo B, Consumo D-4, Consumo D-5, Consumo D-6, Consumo Anual, Consumo Semanal, Temperatura Média, Temperatura Máxima e a Velocidade Média do Vento encontram-se ilustradas nos gráficos 4.20 a 4.27.

Para cada uma destas variáveis que não foram inseridas no modelo B, a representação gráfica dos resíduos não apresenta uma relação entre os resíduos e cada uma destas variáveis. Assim, conclui-se uma vez mais que estas variáveis não têm influência na variável resposta.

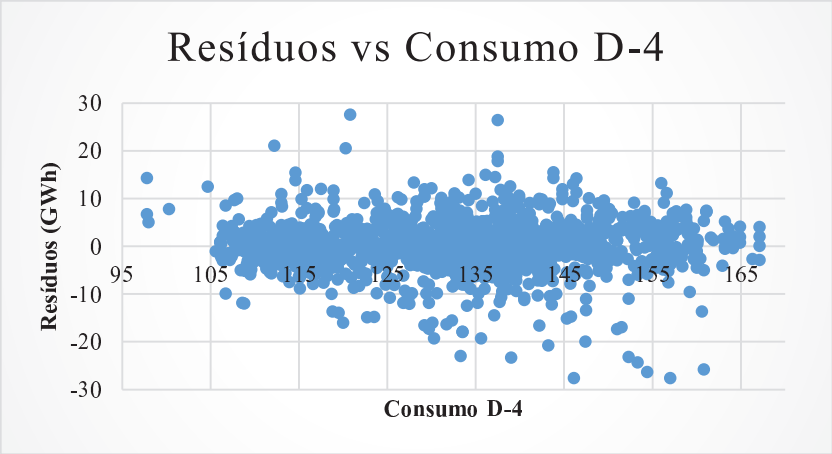

Gráfico 4.20: Representação gráfica dos Resíduos do modelo B contra a variável Consumo D-4

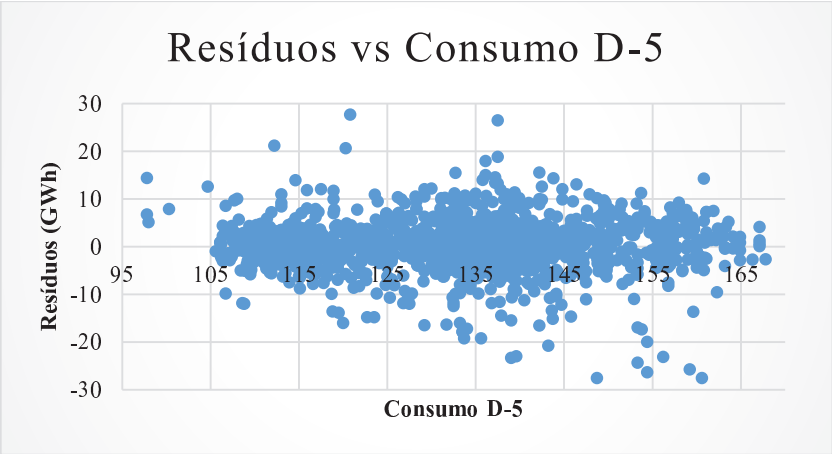

Gráfico 4.21: Representação gráfica dos Resíduos do modelo B contra a variável Consumo D-5

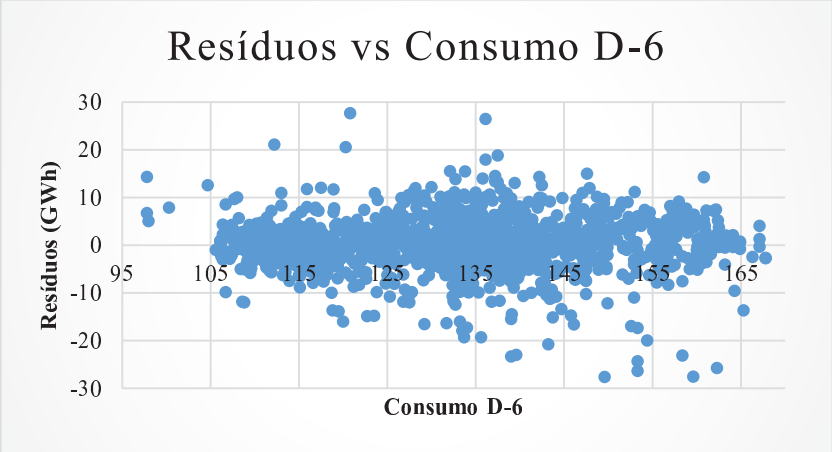

Gráfico 4.22: Representação gráfica dos Resíduos do modelo B contra a variável Consumo D-6

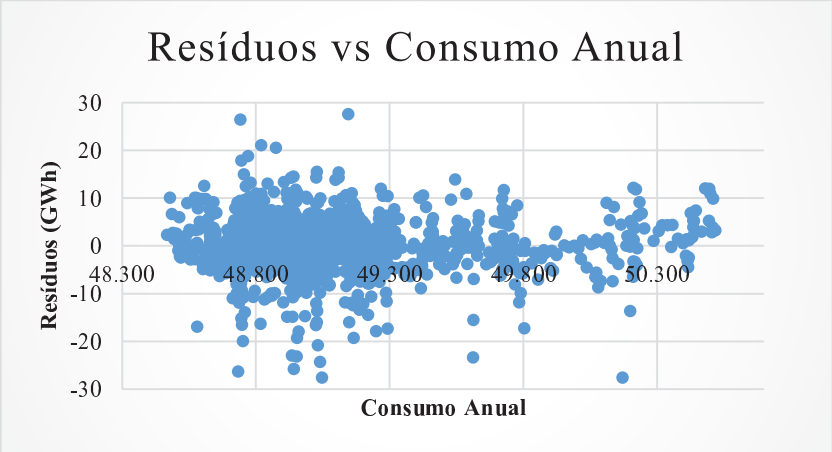

Gráfico 4.23: Representação gráfica dos Resíduos do modelo B contra a variável Consumo Anual de Energia Elétrica

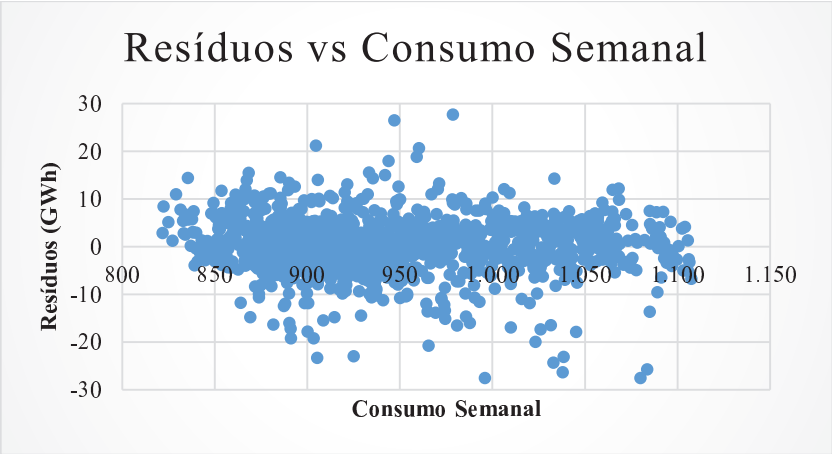

Gráfico 4.24: Representação gráfica dos Resíduos do modelo B contra a variável Consumo Semanal de Energia Elétrica

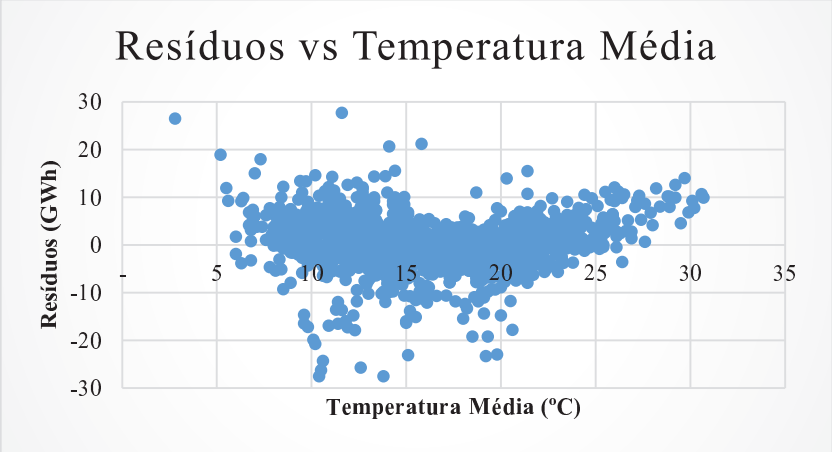

Gráfico 4.25: Representação gráfica dos Resíduos do modelo B contra a variável Temperatura Média Diária em Portugal

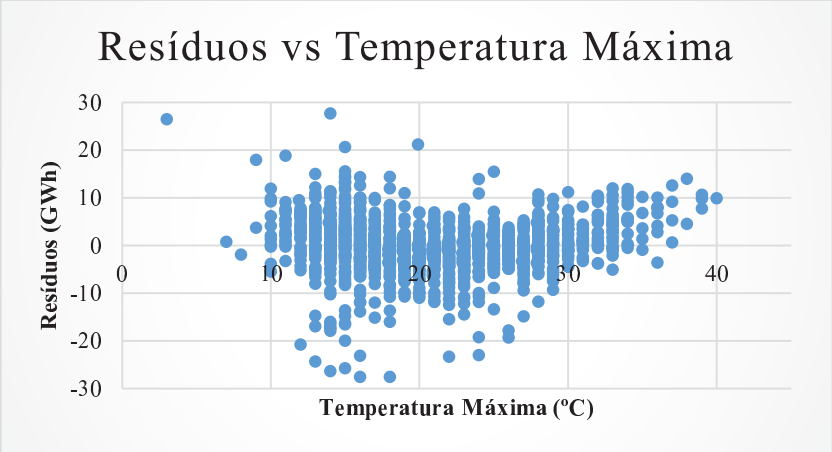

Gráfico 4.26: Representação gráfica dos Resíduos do modelo B contra a variável Temperatura Máxima Diária em Portugal

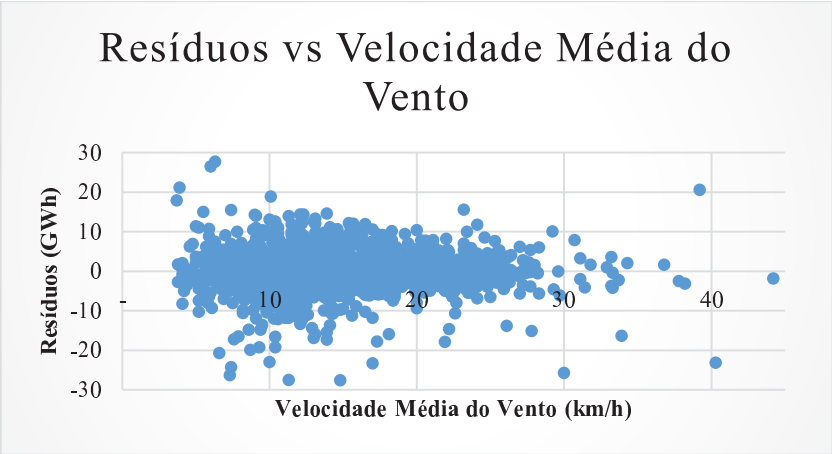

Gráfico 4.27: Representação gráfica dos Resíduos do modelo B contra a variável Velocidade Média do Vento diária em Portugal

A representação gráfica dos resíduos contra os valores ajustados do modelo B encontra-se descrita no gráfico 4.28. Através deste gráfico, conclui-se que não é necessário executar nenhuma modificação às variáveis incluídas no modelo de regressão linear múltipla B e que existe um bom ajustamento neste modelo, uma vez que a incidência dos valores ajustados é para resíduos próximos de zero.

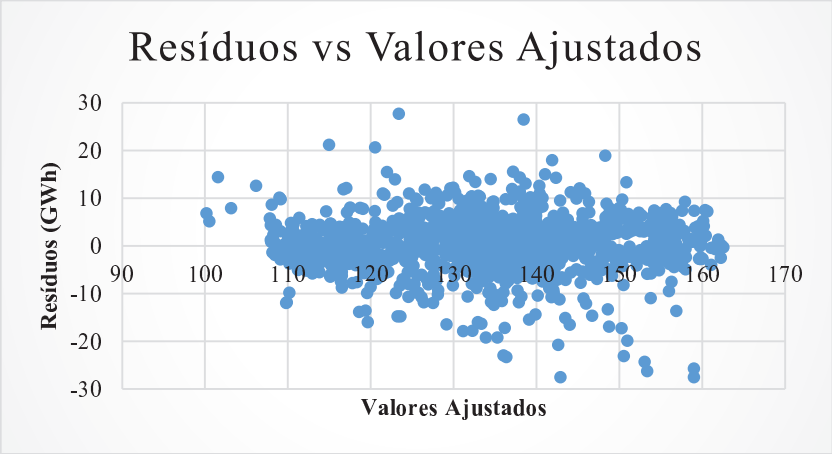

Gráfico 4.28: Representação gráfica dos Resíduos contra os Valores Ajustados do modelo B

Por fim, o estudo da normalidade dos resíduos para o modelo B foi realizado através do Teste de *Lilliefors,* em que as hipóteses de teste são:

H_0: Os resíduos seguem uma distribuição normal *versus* H_1:Os resíduos não seguem uma distribuição normal

Os resíduos do modelo de regressão múltipla B contêm as seguintes caraterísticas amostrais, apresentadas na tabela 4.14:

Tabela 4.14: Principais caraterísticas amostrais do modelo de regressão múltipla B

Caraterísticas amostrais do Modelo B					
Tamanho da Amostra, n	Média	Desvio Padrão	Nível de Significância, α	d	$dc_{n,\propto}$
1461	0	5,46	0,05	1,00	0,02

Como $d > dc_{1461,0.05}$, a hipótese nula é rejeitada, ou seja, rejeita-se a hipótese de os resíduos seguirem uma distribuição normal.

A construção do histograma para os resíduos do modelo B reforçam a ideia de que estes não seguem uma distribuição normal, como se poderá confirmar pelo gráfico 4.29, em que valores extremos dos resíduos e os valores médios não se ajustam à distribuição normal.

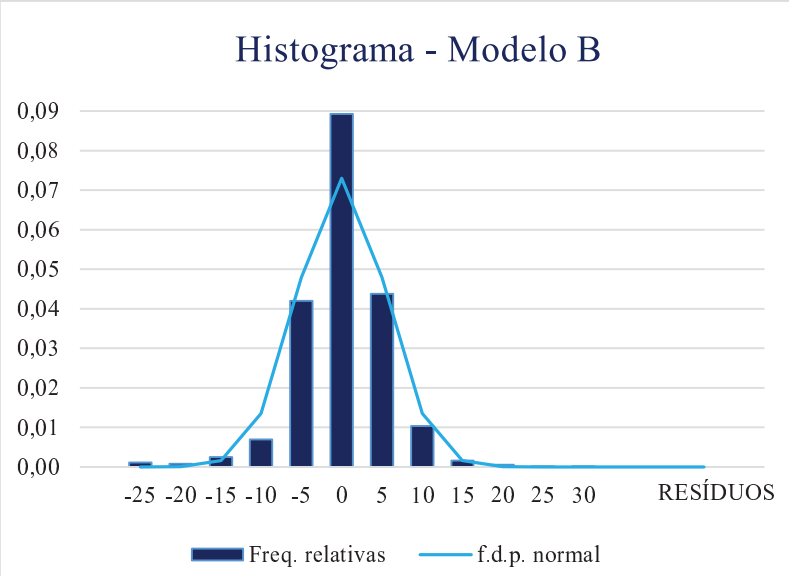

Gráfico 4.29: Histograma dos Resíduos do Modelo B e a função distribuição de probabilidade da Normal

4.5.3 Análise dos Resíduos do Modelo C

O modelo C é formado pelas variáveis independentes Consumo D-3, Consumo D-7, Consumo Mensal, Temperatura Mínima diária, Temperatura Média Diária e Temperatura Máxima Diária. A representação gráfica 4.30 a 4.35 nega a existência de uma relação entre os resíduos e cada uma das variáveis inseridas no modelo C.

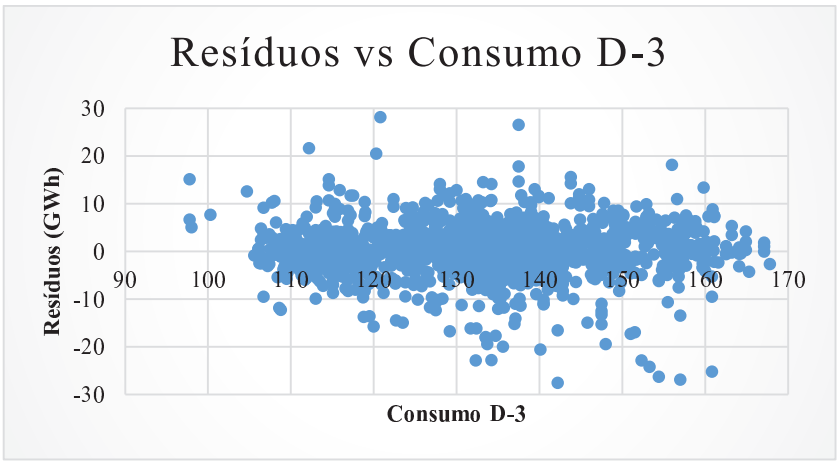

Gráfico 4.30: Representação gráfica dos Resíduos do Modelo C *versus* a variável de Consumos de Energia Elétrica em Portugal de há 3 dias atrás

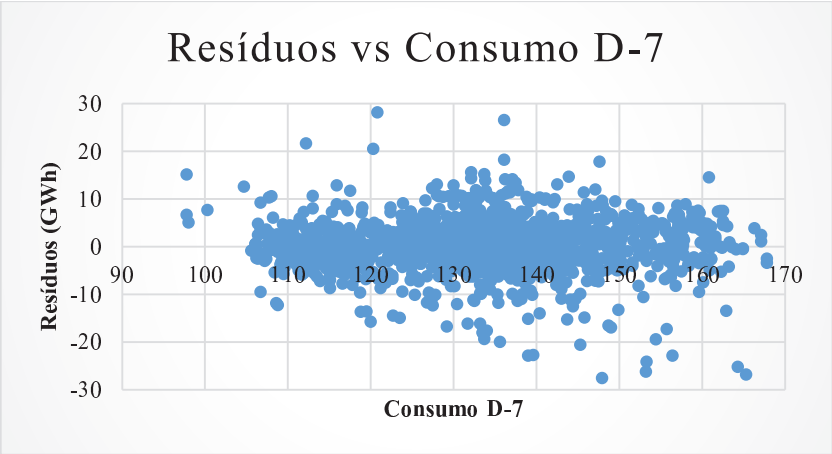

Gráfico 4.31: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo D-7

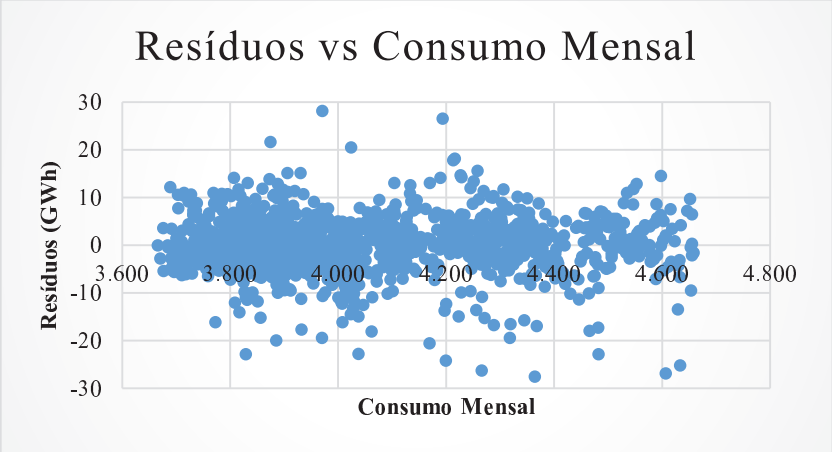

Gráfico 4.32: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo Mensal de Energia Elétrica em Portugal

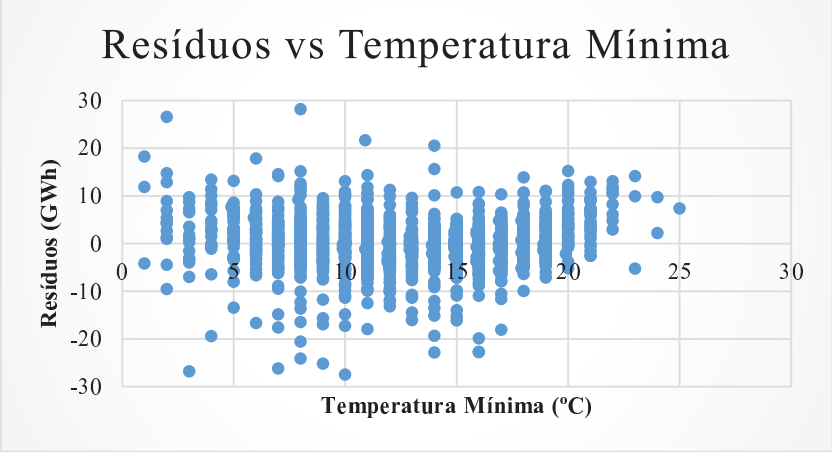

Gráfico 4.33: Representação Gráfica dos Resíduos do Modelo C contra a variável Temperatura Mínima Diária em Portugal

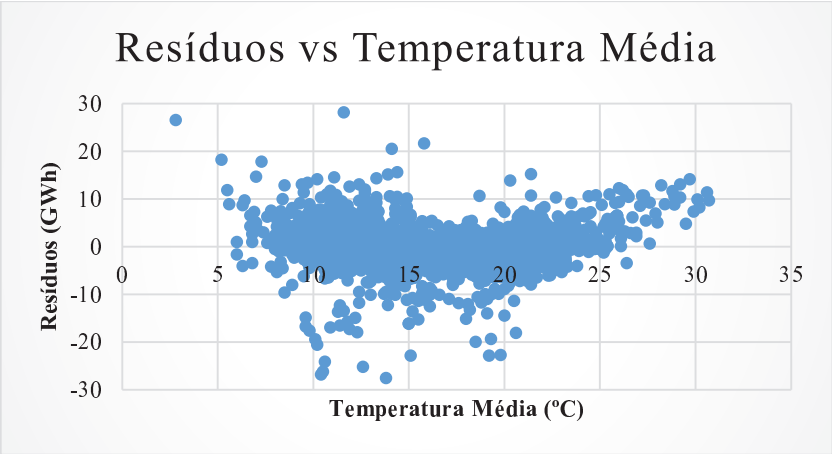

Gráfico 4.34: Representação Gráfica dos Resíduos do Modelo C contra a variável Temperatura Média Diária em Portugal

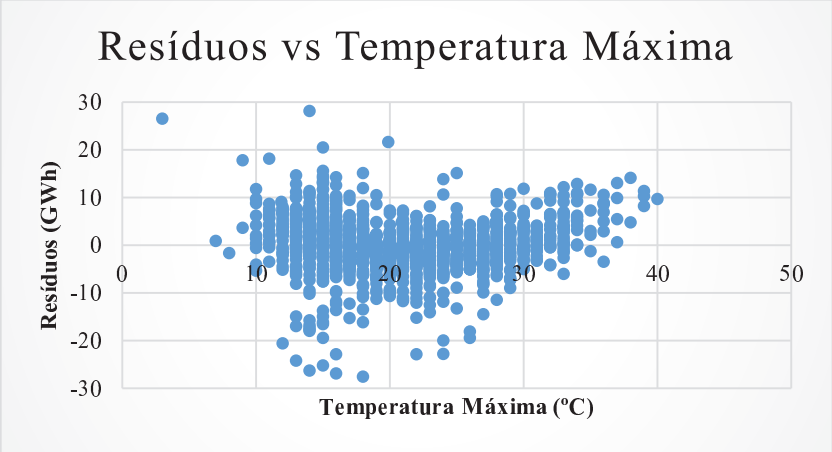

Gráfico 4.35: Representação Gráfica dos Resíduos do Modelo C contra a variável Temperatura Máxima Diária em Portugal

A representação gráfica dos resíduos contra as restantes variáveis independentes que não foram incluídas no modelo C, Consumo D-4, Consumo D-5, Consumo D-6, Consumo Anual, Consumo Semanal e Velocidade Média do Vento encontra-se descrita nos gráficos 4.36 a 4.41.

Através destes gráficos, é evidenciado o facto de não existir uma relação entre os resíduos do modelo C e as variáveis independentes que não foram introduzidas no modelo. Uma vez que não existe uma relação entre os resíduos e nenhuma das variáveis não incluídas no modelo, não é necessário transformar nenhuma destas variáveis independentes.

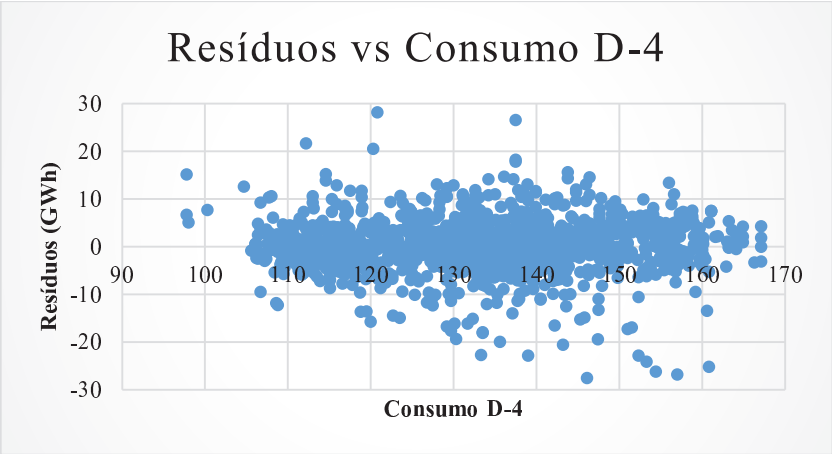

Gráfico 4.36: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo de há 4 dias atrás

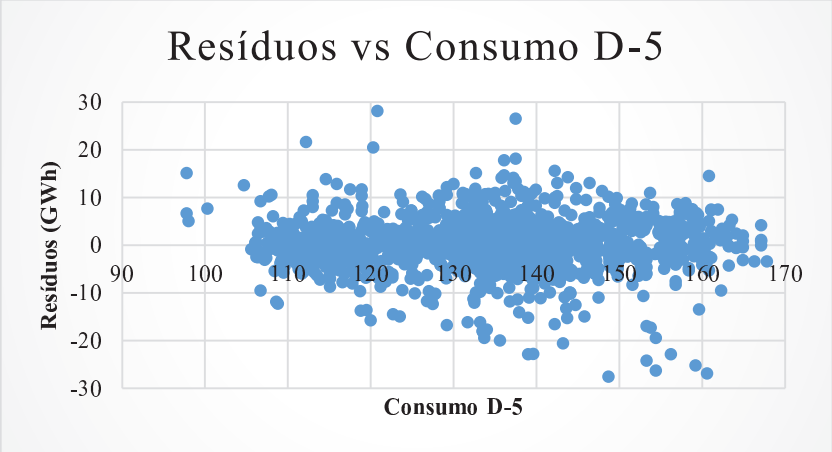

Gráfico 4.37: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo D-5

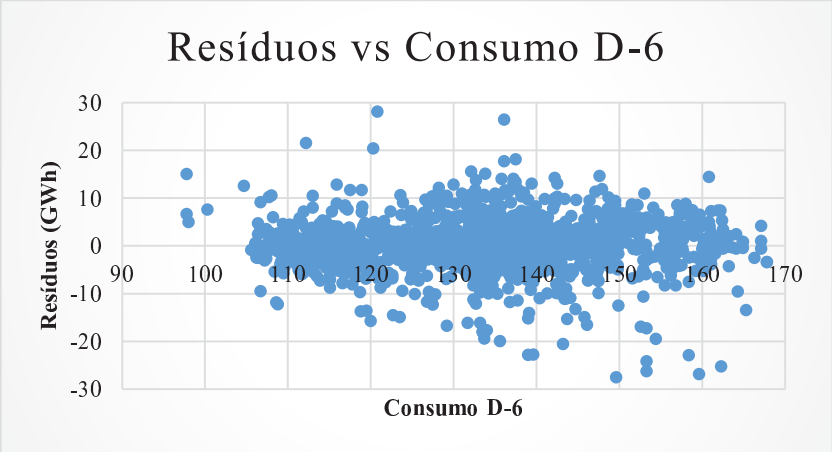

Gráfico 4.38: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo D-6

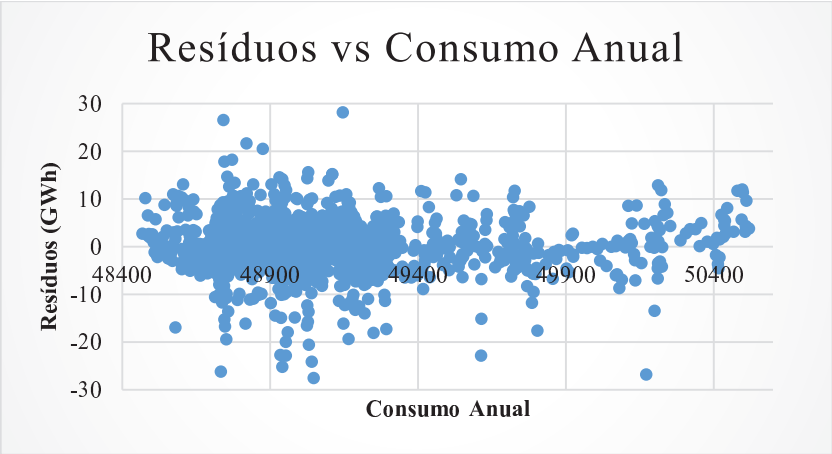

Gráfico 4.39: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo Anual de Energia Elétrica Em Portugal

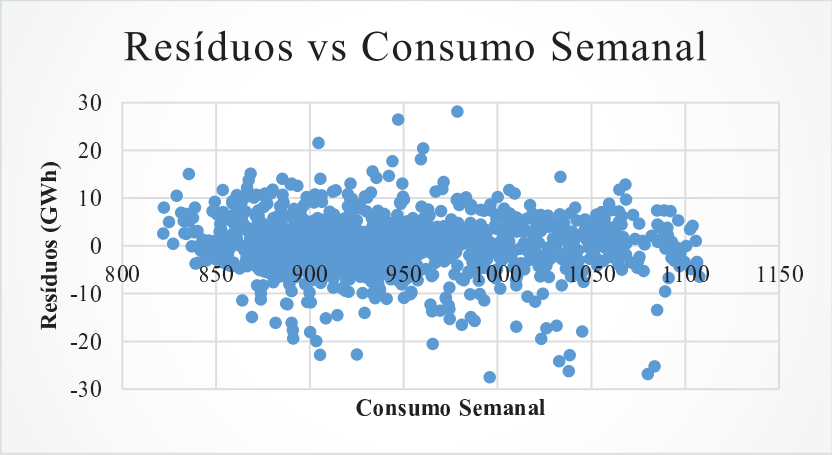

Gráfico 4.40: Representação Gráfica dos Resíduos do Modelo C contra a variável Consumo Semanal de Energia Elétrica

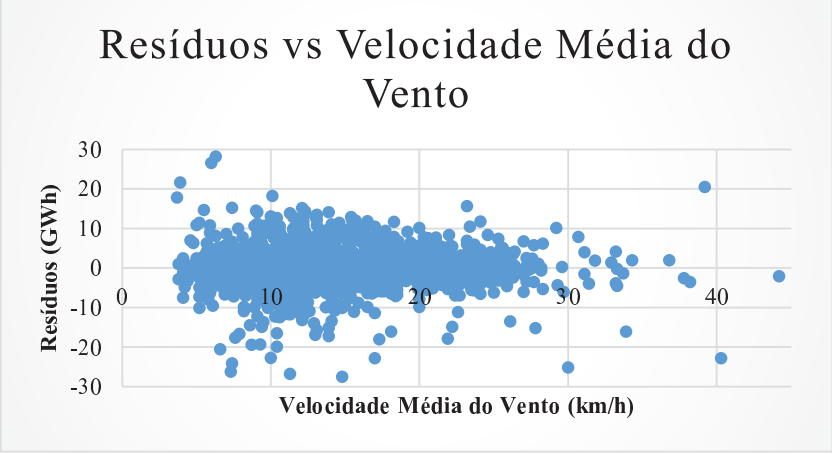

Gráfico 4.41: Representação Gráfica dos Resíduos do Modelo C contra a variável Velocidade Média Diária do Vento em Portugal

A representação gráfica dos resíduos contra os valores ajustados do modelo C encontra-se descrita no gráfico 4.42. Através deste gráfico, conclui-se que não é necessário executar nenhuma modificação às variáveis incluídas no modelo de regressão linear múltipla C.

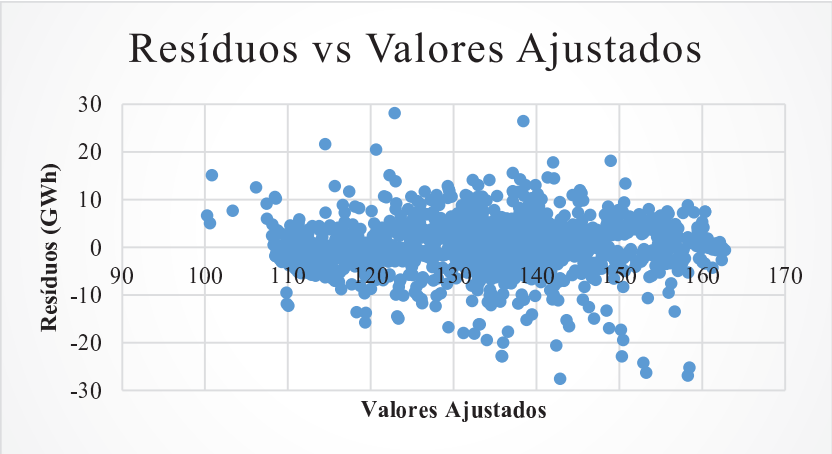

Gráfico 4.42: Representação Gráfica dos Resíduos *versus* Valores Ajustados do Modelo C

Por fim, o estudo da normalidade dos resíduos foi realizado com o Teste de *Liliefors* e, deste modo, as hipóteses de teste são:

H_0: Os resíduos seguem uma distribuição normal *versus* H_1:Os resíduos não seguem uma distribuição normal

Os resíduos do modelo C são definidos pelas seguintes caraterísticas amostrais, definidas na tabela 4.15.

Tabela 4.15: Principais caraterísticas amostrais do modelo de regressão múltipla C

Caraterísticas amostrais do Modelo C					
Tamanho da Amostra, *n*	Média	Desvio Padrão	Nível de Significância, α	d	$dc_{n,\propto}$
1461	0	5,47	0,05	1,00	0,02

Como $d > dc_{1461,0.05}$, a hipótese nula é rejeitada, ou seja, rejeita-se a hipótese de os resíduos seguirem uma distribuição normal.

Esta conclusão é reforçada com a construção do histograma para os resíduos do modelo C, como se poderá visualizar no gráfico 4.43, uma vez que não há uma boa adaptação dos resíduos à função distribuição da Normal.

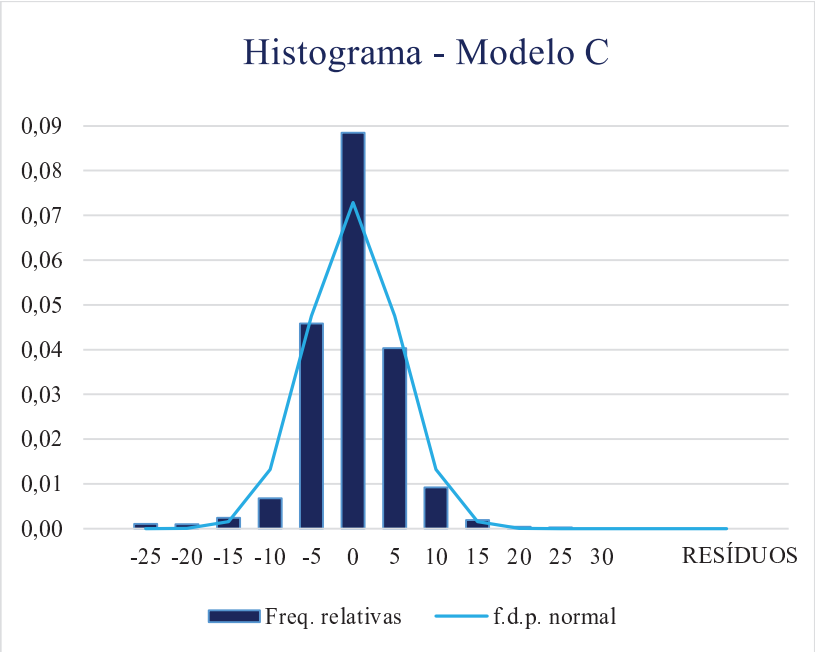

Gráfico 4.43: Histograma dos Resíduos do Modelo C contra a função de distribuição da Normal

É necessário realçar novamente a existência de 3 variáveis independentes incluídas no modelo de RLM C com multicolinearidade, Temperatura Mínima, Média e Máxima diária.

Apesar da normalidade dos resíduos ser rejeitada para qualquer um dos três modelos de regressão múltipla, através do teorema da máxima verosimilhança temos a garantia da normalidade e "optimalidade" assintótica dos estimadores, mesmo para distribuições não normais.

4.6 Intervalos de Predição

Construídos os modelos de regressão linear múltipla que melhor descrevem os consumos de energia elétrica em Portugal é tempo de perceber qual dos 3 modelos melhores resultados fornece.

Anteriormente foram construídos 3 modelos de regressão linear múltiplos, A, B e C com métodos de seleção de variáveis distintos e partindo de diferentes conjuntos de variáveis independentes. Já foi também referido a existência de problemas de multicolinearidade no modelo C, fazendo deste modelo o menos apto para a previsão de consumos de energia elétrica em Portugal.

Posto isto, torna-se necessário perceber qual dos 3 modelos tem melhores resultados ao nível de predição.

A escolha dos valores não observados de variáveis independentes foi feita tendo em conta os extremos das variáveis inseridas nos modelos, onde a matriz X^* é constituída pela combinação dos valores máximos, médios e mínimos de cada variável. Este pressuposto foi assumido com o objetivo de se criarem cenários em situações extremas, ou seja, de se criarem os piores cenários possíveis.

Relembrando o que foi referido na secção 3.2.4, a predição de $\hat{y}^*$ é dada pela equação 4.7 e o intervalo de $(1-\propto) \times 100\%$ de confiança para y^* é definido pela expressão 4.8, em que $t_{n-m}^{1-\frac{\propto}{2}}$ representa o quantil de ordem $1-\frac{\propto}{2}$ da distribuição *t-student* com *n-m* graus de liberdade.

$$\hat{y}^* = \sum_{j=1}^{m} x_j^* \hat{b}_j = X^{*T} \hat{b}$$

(4.7)

$$\left(\hat{y}^* - t_{n-m}^{1-\frac{\propto}{2}} S \sqrt{X^{*T} \left(X^T X \right)^{-1} X^* + 1};\ \hat{y}^* + t_{n-m}^{1-\frac{\propto}{2}} S \sqrt{X^{*T} \left(X^T X \right)^{-1} X^* + 1} \right)$$

(4.8)

4.6.1 Modelo A

O modelo de regressão linear múltipla A é composto apenas por variáveis internas ao consumo de energia elétrica em Portugal e a predição de $\hat{y}^*$ para o modelo A é dada por:

$$\hat{y}^* = X^{*T} \hat{b} = [1 \quad x_{Consumo\ D-3} \quad x_{Consumo\ D-7} \quad x_{Consumo\ Mensal}] \begin{bmatrix} 20{,}09 \\ 0{,}57 \\ 0{,}41 \\ -4{,}04\text{E}-04 \end{bmatrix}$$

(4.9)

Como é de esperar, a amplitude do intervalo de predição depende sempre da escolha dos valores para as variáveis independentes Consumo D-3, Consumo D-7 e Consumo Mensal. Tal como foi mencionado na secção 4.6, os valores usados na construção dos intervalos de predição correspondem à combinação dos máximos, médias e mínimos das respetivas variáveis, existindo assim 27 intervalos possíveis nos piores cenários possíveis, como por exemplo um dia com o consumo de há 3 dias muito baixo, o consumo de há 7 dias muito alto e o consumo mensal muito alto.

Construídos os 27 cenários possíveis, o tamanho do intervalo de predição a nível de 95% de confiança mais pequeno foi de 0,015, enquanto que o maior foi de 2,766.

Uma vez que a amplitude dos intervalos de predição é reduzida nos piores cenários criados, espera-se que em cenários mais "realistas", a amplitude seja inferior ao do maior intervalo obtido e, deste modo, conclui-se que o modelo A é um bom modelo de previsão de consumos de energia elétrica em Portugal.

4.6.2 Modelo B

O modelo de regressão linear múltipla B é composto por variáveis internas e externas ao consumo de energia elétrica em Portugal.

A predição de $\hat{y}^*$ para o modelo B é dada por:

$$\hat{y}^* = X^{*T} \hat{b} = [1 \quad x_{Consumo\ D-3} \quad x_{Consumo\ D-7} \quad x_{Consumo\ Mensal} \quad x_{Temperatura\ Mínima}] \begin{bmatrix} 34{,}17 \\ 0{,}57 \\ 0{,}41 \\ -0{,}01 \\ -0{,}23 \end{bmatrix}$$

(4.10)

Como é de esperar, a amplitude do intervalo de predição depende sempre da escolha dos valores para as variáveis independentes Consumo D-3, Consumo D-7, Consumo Mensal e Temperatura Mínima. Tal como foi mencionado na secção 4.6, os valores usados na construção dos intervalos de predição correspondem à combinação dos máximos, médias e mínimos das respetivas variáveis, existindo assim 81 intervalos nos piores cenários possíveis, por exemplo, um dia com o consumo de há 3 dias muito baixo, o consumo de há 7 dias muito alto, o consumo mensal muito alto e a temperatura mínima muito elevada.

Construídos os 81 cenários, o tamanho do intervalo de predição ao nível de 95% de confiança mais pequeno foi de aproximadamente 0, enquanto que o maior foi de 2,9.

Deste modo, considera-se que o modelo B é um bom modelo para prever os consumos de energia elétrica em Portugal, uma vez que as amplitudes dos intervalos nos piores cenários possíveis são pequenas é de se esperar que em cenários mais "realistas" esses valores sejam inferiores ao do maior tamanho do intervalo obtido.

4.7.3 Modelo C

O modelo de regressão linear múltipla C é composto por variáveis internas e externas ao consumo de energia elétrica em Portugal e a predição de $\hat{y}^*$ para este modelo é dada por:

$$\hat{y}^* = X^{*T}\hat{b}$$

$$= [1 \quad x_{Cons.\ D-3} \quad x_{Cons.\ D-7} \quad x_{Cons.\ Mensal} \quad x_{Temp.\ Mín} \quad x_{Temp.\ Med} \quad x_{Temp.\ Max}] \begin{bmatrix} 34{,}04 \\ 0{,}56 \\ 0{,}41 \\ -0{,}01 \\ -0{,}57 \\ 0{,}69 \\ -0{,}33 \end{bmatrix} \tag{4.11}$$

Tal como foi referido anteriormente, a amplitude do intervalo de predição depende sempre da escolha dos valores para as variáveis independentes incluídas no modelo C: Consumo D-3, Consumo D-7, Consumo Mensal, Temperatura Mínima, Temperatura Média e Temperatura Máxima. Tal como foi mencionado na secção 4.6, os valores usados na construção dos intervalos de predição correspondem à combinação dos máximos, médias e mínimos das respetivas variáveis, existindo assim 729 intervalos nos piores cenários possíveis, por exemplo, um dia com o consumo de há 3 dias muito baixo, o consumo de há 7 dias muito alto, o consumo mensal muito alto, a temperatura mínima muito elevada, a temperatura média muito baixa e a temperatura máxima muito alta.

Construídos os 729 cenários, o tamanho do intervalo de predição a nível de 95% de confiança mais pequeno foi de aproximadamente 0, enquanto o maior foi de 43,7.

Como já foi referido anteriormente, não é aconselhável a utilização do modelo C para prever os consumos de energia elétrica em Portugal, uma vez que existe multicolinearidade em 3 das variáveis que o compõem. Este facto é evidenciado mais uma vez pela amplitude dos intervalos de predição testados. Deste modo, é novamente reforçada a ideia de que este modelo não é um bom modelo de RLM para descrever os consumos de energia elétrica em Portugal.

4.7 Teste dos Modelos

Nesta secção pretende-se demonstrar os resultados apenas para os modelos A e B para dados referentes a janeiro e fevereiro de 2016. É de salientar que estes dados não fazem parte dos consumos de energia elétrica e temperatura utilizadas na criação dos modelos.

A evolução do consumo real de eletricidade em Portugal e a previsão dos consumos de energia elétrica através do modelo A encontra-se descrita no gráfico 4.44.

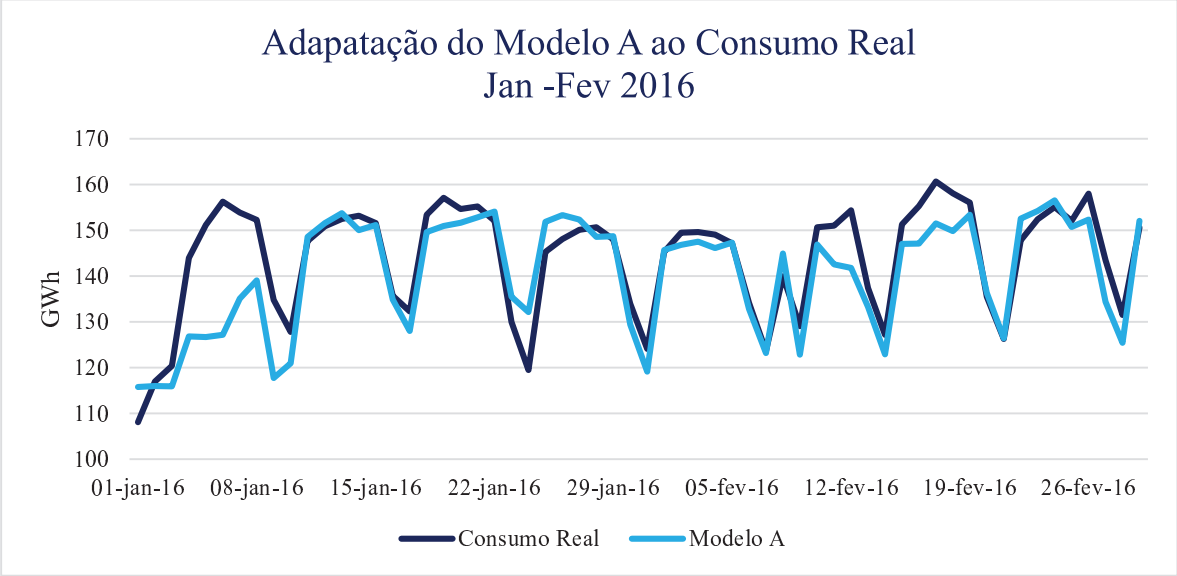

Gráfico 4.44: Representação gráfica do consumo real e do consumo de energia elétrica obtido através do modelo A em Portugal, desde janeiro a fevereiro de 2016

A evolução do consumo real de eletricidade em Portugal e a previsão dos consumos de energia elétrica através do modelo B encontra-se descrita no gráfico 4.45:

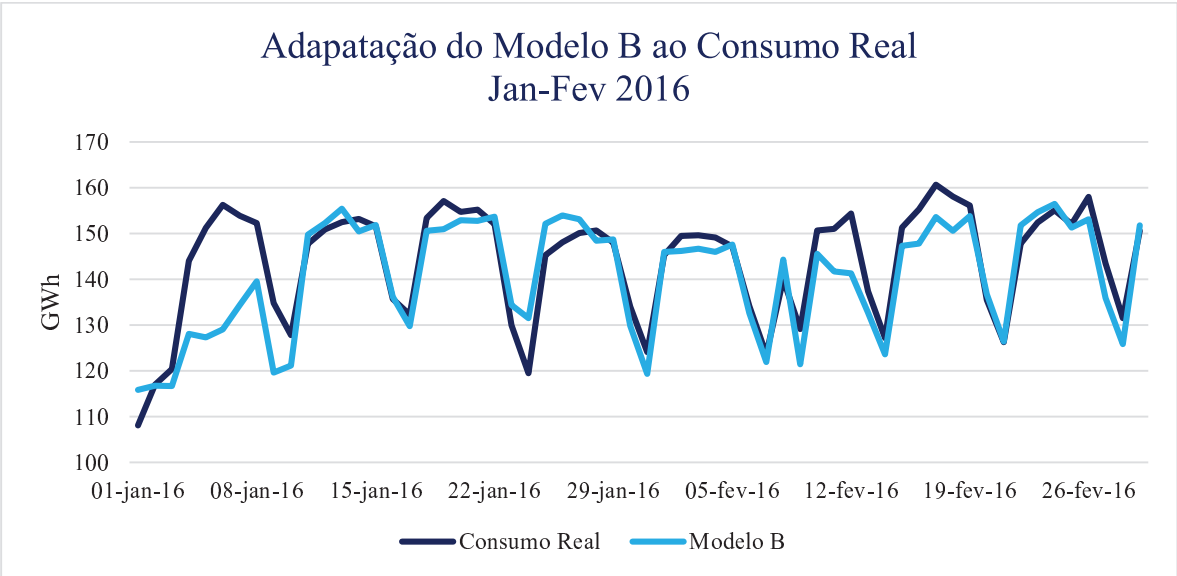

Gráfico 4.45: Representação gráfica do consumo real e do consumo de energia elétrica obtido através do modelo B em Portugal, desde janeiro a fevereiro de 2016

Visualmente, os resultados obtidos pelo modelo A e o modelo B são semelhantes. É de notar o mau ajustamento dos modelos na primeira semana de janeiro e na semana com o Carnaval em fevereiro. Este mau ajustamento deve-se à sazonalidade caraterística da época festiva.

O método estatístico utilizado para avaliar qual dos modelos obtém os melhores resultados encontra-se descrito na expressão 4.12, em que n representa o tamanho da amostra e e_i os resíduos do modelo a ser testado. Quanto mais pequeno for o valor da expressão 4.12, melhores resultados se obtém dos modelos de regressão múltipla.

$$\frac{\sum_{i=1}^{n}(e_i)^2}{n} \tag{4.12}$$

Os resultados do método estatístico apresentado na expressão 4.12 indicam que o modelo B apresenta melhores resultados na previsão de consumos de energia elétrica em Portugal. Estes resultados encontram-se descritos na tabela 4.16.

Tabela 4.16: Avaliação dos modelos segundo o método estatístico definido em Expressão 4.12

Modelo A	Modelo B
65,00	59,96

Deste modo, o modelo que melhor representa a previsão de consumos de energia elétrica em Portugal é o modelo de regressão linear múltipla B. Note-se ainda que graficamente o modelo B poderia ter um ajustamento melhor face ao consumo real de janeiro e fevereiro de 2016.

Considerações finais e problemas em aberto

Os modelos de previsão de consumos diários de energia elétrica a curto prazo, consideram o consumo num determinado dia *t* como sendo uma combinação linear de alguns dias anteriores, ou seja, variáveis dependentes são também consideradas variáveis explicativas. Do ponto de vista estatístico, surge assim um modelo condicional a consumos de energia elétrica de dias anteriores que pode, até certo ponto, ser tratado como um modelo de regressão linear múltipla.

A variável resposta que se pretende obter com os modelos criados é uma boa previsão de consumos de energia elétrica em Portugal diariamente. Com isto, as variáveis explicativas a serem analisadas foram caraterizadas de maneiras distintas: variáveis internas e variáveis externas ao consumo.

A criação dos modelos de regressão linear múltipla foi feita tendo em conta vários pressupostos. Os métodos de seleção de variáveis utilizados foram a Regressiva e o *Stepwise,* utilizando variáveis históricas de consumos (variáveis internas) e/ou a variáveis externas ao consumo de energia elétrica tais como a temperatura e a velocidade do vento.

Os modelos de regressão linear múltipla obtidos neste relatório apresentam o mesmo nível de ajustamento face aos valores observados, quer as variáveis explicativas sejam dependentes do consumo ou sejam consideradas também variáveis externas a este.

A escolha entre os modelos de regressão linear múltipla obtidos neste estudo, modelo A e o modelo B, passa por perceber se faz sentido a introdução de uma variável externa ao consumo de energia elétrica em Portugal, a Temperatura Mínima Diária. Deste modo, é necessário perceber se faz sentido ou não a introdução desta variável externa ao consumo, uma vez que se estará dependente de informação de terceiros.

É de salientar que ainda que o modelo de regressão linear múltipla B apresente resultados ligeiramente melhores, o uso do modelo A devolve resultados "equivalentes" a este, tendo a vantagem de estar apenas dependente de informação interna ao consumo de energia elétrica em dias anteriores.

Como já foi referido, os modelos de regressão múltipla obtidos têm espaço para melhorias como, por exemplo, a inclusão de mais informação sobre se o consumo se refere a um dia de época festiva.

Em particular, faria sentido testar a introdução de outras variáveis externas ao consumo de energia elétrica em Portugal tais como o nível de precipitação, o número diário de horas com luz solar, entre outras, uma vez que estas variáveis influenciam diretamente o consumo de energia elétrica no dia-a-dia nas nossas instalações/habitações.

Posteriormente, seria interessante testar modelos de RLM para cada nível de tensão, uma vez que estes modelos apenas refletem a previsão de consumos de energia elétrica de modo agregado. Como cada nível de tensão tem a sua peculiaridade é de se esperar que o ajustamento dos modelos de regressão linear obtidos seja melhorado ao se realizar a previsão de consumos por nível de tensão e não como um todo. Deste modo, faria sentido testar modelos de regressão linear múltipla para cada um dos 5 níveis de tensão e, consequentemente, testar a introdução das variáveis referidas anteriormente para teste.

Referência Bibliográfica

Alpuim, T., Modelos Lineares – Notas de apoio à disciplina, 2014

Bachir, N., (2015). Fatores Explicativos na Oferta de Serviços Farmacêuticos em Portugal. Tese de Mestrado em Matemática Aplicada à Economia e Gestão – Faculdade de Ciências da Universidade de Lisboa.

Casella, G. e Berger, R.L. (2002). *Statistical Inference*. 2ª edição, *Duxbury advanced series*

EDP (2012). O que é energia elétrica. Acedido em: 10 de julho de 2016, em:

http://www.edp.com.br/pesquisadores-estudantes/energia-eletrica/o-que-e-energia-eletrica/Paginas/default.aspx#6

EDP (2009). Sistema Elétrico Português. Acedido em: 10 de julho de 2016, em:

http://www.edp.pt/pt/aedp/sectordeenergia/sistemaelectricoportugues/Pages/SistElectNacional.aspx

ERSE (2016, janeiro). Guia de Medição, Leitura e Disponibilização de Dados. Acedido em: 9 de agosto de 2016, em:

www.erse.pt/pt/electricidade/regulamentos/relacoescomerciais/Documents/SubRegulamentação/GMLDD_2016.pdf

EURONEWS. Previsão Meteorológica por Região. Acedido em: 6 de março de 2016, em:

http://pt.euronews.com/meteorologia/europa/portugal

História da Eletricidade. Acedido em: 7 de agosto de 2016, em:

http://www.sofisica.com.br/conteudos/HistoriaDaFisica/historiadaeletricidade.php

REN (2015). O setor elétrico. Acedido em: 10 de julho de 2016, em:

https://www.ren.pt/pt-PT/o_que_fazemos/eletricidade/o_setor_eletrico/#1

REN (2012). Relatório e Contas 2011. Acedido em: 8 de agosto de 2016, em:

http://relatorioecontas2011.ren.pt

REN (2013). Relatório e Contas 2012. Acedido em: 8 de agosto de 2016, em:

http://relatorioecontas2012.ren.pt

REN (2014). Relatório e Contas 2013. Acedido em: 8 de agosto de 2016, em:

http://relatorioecontas2013.ren.pt

REN (2015). Relatório e Contas 2014. Acedido em: 8 de agosto de 2016, em:

http://relatorioecontas2014.ren.pt

REN (2016). Relatório e Contas 2015. Acedido em: 8 de agosto de 2016, em:

http://relatorioecontas2015.ren.pt

REN. Estatística Diária – SEN. Acedido em: 5 de março de 2016, em:

http://www.centrodeinformacao.ren.pt/PT/Paginas/CIHomePage.aspx

Tempo em Lisboa. Acedido em: 6 de março de 2016, em:

http://www.tempo.pt/historico/lisboa.htm

Universidade do Minho (2007-2008). Tabelas Estatísticas. Acedido em: 27 de agosto de 2016, em:

http://pessoais.dps.uminho.pt/lac/2007-2008/EST2/tabelas_estatisticas.pdf

Young, G.G e Smith, R.L. (2005). *Essentials of Statistical Inference*. *Cambridge University Press*

Anexos

Anexo 1 – Tabela de Quantis para a Estatística de Teste de *Lilliefors* para a Distribuição Normal

Quantis para a Estatística de Teste de Lilliefors para a Distribuição Normal

p =	*0.80*	*0.85*	*0.90*	*0.95*	*0.99*
n= 4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.142	.147	.158	.173	.200
30	.131	.136	.144	.161	.187
> 30	$\frac{.736}{\sqrt{n}}$	$\frac{.768}{\sqrt{n}}$	$\frac{.805}{\sqrt{n}}$	$\frac{.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

Fonte: tabela adaptada de Lilliefors (1967)