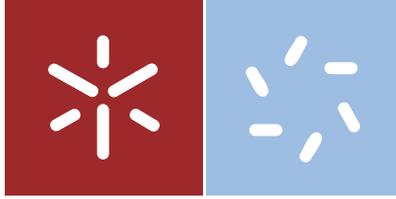


Universidade do Minho
Escola de Ciências

Luís Paulo Ferreira da Costa

Modelação de dados do serviço de
urgência no Hospital de Braga



Universidade do Minho
Escola de Ciências

Luís Paulo Ferreira da Costa

Modelação de dados do serviço de
urgência no Hospital de Braga

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação do
Professora Doutora Inês Sousa
Doutora Sónia Dória

Declaração

Nome: Luís Paulo Ferreira da Costa

Endereço eletrónico: luispfcosta@gmail.com

Título da dissertação:

Modelação de dados do serviço de urgência no Hospital de Braga

Orientadores:

Professora Doutora Inês Sousa

Doutora Sónia Dória

Ano de conclusão: 2018

Mestrado em Estatística

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA DISSERTAÇÃO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 31 de Janeiro de 2018.

O autor: Luís Costa

AGRADECIMENTOS

Começo por agradecer à minha orientadora Inês Sousa que ao longo deste trabalho sempre me apoiou e sempre esteve disponível.

Outro elemento muito importante para que esta dissertação fosse possível ser realizada é o Hospital de Braga, agradeço a oportunidade de fazer parte deste projecto.

Queria dar os meus agradecimentos a todos os meus amigos, uma vez que foram essenciais para a conclusão desta dissertação. Não quero dizer nomes, pois eles sabem quem são, mas em todo o caso aqui vão eles: Jorge, Partyboy, Cris, HD, Susana, Cláudio... Foram pessoas que sempre deram aquela vontade que me faltava para escrever.

Para último deixo todos os meus agradecimento aos meus pais, pelo apoio incondicional que sempre me deram ao longo deste trabalho.

Obrigado a todos!!

ABSTRACT

This work is about the capacity of forecasting the daily number of patients that enter Hospital of Braga Emergency Room (HBER). For this study we have annual time series of the daily number of patients since 2012 up to 2016, and having that we then do multiple regression analysis with correlated errors. In that analysis we identify the following problems: (1) the errors of the model are not independent, (2) some of the explanatory variables are not significant, (3) the model is not considering the errors of an autoregressive process.

The errors of an autoregressive model are going to be considered has following an ARIMA process. We defined two regression models in which the error term of both models, follow an autoregressive process of order seven (AR(7)).

In the first model the forecasting is done for a long time period (one month or one year), in the second model we use the daily observed data, in real time, so that the forecasting is altered in relation to the first model, doing that we can obtain better forecasts.

Comparing the forecasts, for the same period, we can observe that the mean absolute percentage error from the first model is higher than the second model.

With the help of R statistical software, more precisely the *Shiny package*, we were able to build a forecasting Web-App. Using this App we can forecast using both of the models that we study in this paper.

Keywords: Multiple Linear Regression, Autocorrelated Errors, Forecasting, Shiny App.

RESUMO

Este trabalho consiste num estudo sobre a capacidade de previsão do número diário de pacientes que são admitidos no serviço de urgências do Hospital de Braga (SUHB). Para este estudo temos ao nosso dispôr séries temporais anuais do número de pacientes, desde 2012 até 2016, sendo feita posteriormente uma análise de regressão linear múltipla com erros correlacionados. Nessa análise são identificados os seguintes problemas: (1) os erros do modelo não são independentes, (2) algumas das variáveis explicativas não são significativas, (3) o termo de erro do modelo é definido por um processo autoregressivo.

Os erros de um modelo autoregressivo irão ser considerados seguindo um processo autoregressivo integrado e de médias móveis (ARIMA). Definiu-se dois modelos de regressão, nos quais os termos de erro de ambos, seguem um processo autoregressivo de ordem sete (AR(7)).

No primeiro modelo fazem-se previsões a longo prazo (um mês ou um ano), no segundo modelo utilizam-se os dados diários observados em tempo real para alterar as previsões já feitas no primeiro modelo, obtendo-se melhores previsões.

Comparando as previsões dos modelos, realizadas no mesmo período de tempo, é possível observar que, o erro percentual absoluto médio do primeiro modelo é superior ao segundo modelo.

Com a ajuda do *software* R, mais precisamente o *package* Shiny, foi possível construir uma web-app de previsão. Esta aplicação permite fazer previsão utilizando ambos os modelos estudados.

Palavras-Chave: Modelos de Regressão Linear, erros autocorrelacionados, Previsão Diária, Aplicação em Shiny.

CONTEÚDO

1	INTRODUÇÃO	1
2	SERVIÇO DE URGÊNCIA DO HOSPITAL DE BRAGA	3
2.1	Urgência Geral	4
2.1.1	Base de Dados	5
3	REGRESSÃO LINEAR	17
3.1	Notação	17
3.2	Modelos de Regressão Linear	18
3.2.1	Estimação	19
3.2.2	Seleção de Modelo	20
3.2.3	MAPE	22
3.2.4	Diagnóstico	22
3.2.5	Inferências no Modelo de Regressão	24
4	MODELOS SÉRIES TEMPORAIS	27
4.1	Notação	28
4.2	Modelos Estacionários Lineares	30
4.2.1	Modelo White Noise	31
4.2.2	Processos autoregressivos de ordem 1, $AR(1)$	32
4.2.3	Processos autoregressivos de ordem p, $AR(p)$	32
4.2.4	Processos médias móveis de ordem 1, $MA(1)$	33
4.2.5	Processos médias móveis de ordem q, $MA(q)$	33
4.2.6	Processos autoregressivo e médias móveis $ARMA(p,q)$	34
4.3	Modelos Não Estacionários Lineares	34
4.3.1	Trasformação Box-Cox	34
4.3.2	Diferenciação	35
4.4	Modelos de Séries Temporais com Covariáveis	36
4.4.1	Modelos de Regressão Dinâmicos	36
5	MODELAÇÃO DOS DADOS DO SERVIÇO DE URGÊNCIA DO HOSPITAL DE BRAGA	41
5.1	Análise Modelo do SUHB	42
5.1.1	Especificação do Modelo	42
5.1.2	Modelo Final e Previsão	46
6	APLICAÇÃO SHINY PARA URGÊNCIA DO HOSPITAL DE BRAGA	55
6.1	Estrutura de uma aplicação Shiny	56
6.2	Partilhar Aplicação Shiny	59

x **Conteúdo**

6.3	Hospital de Braga: Web-App de Previsão	60
6.3.1	Importação Base de Dados	61
6.3.2	Previsão	62
6.3.3	Estatísticas	64
6.3.4	Publicação Aplicação	65
7	DISCUSSÃO DE RESULTADOS E CONCLUSÕES	67

LISTA DE ABREVIACOES

SUHB	Servio de urgncia do Hospital de Braga
ST	Srie Temporal ou Sries Temporais
AR	Processo autoregressivo
MA	Processo de mdias mveis
ARMA	Processo autoregressivo e mdias mveis
ARIMA	Processo autoregressivo integrado e de mdias mveis
FAC	Funo de autocorrelao
FACP	Funo de autocorrelao parcial
DR	Dynamic Regression
UI	User Interface
HTML	Documento Web

LISTA DE FIGURAS

Figura 1	Diagrama Serviço de Urgências do Hospital de Braga	4
Figura 2	Tempo previsto até à 1 ^o observação médica segundo Triagem de Manchester	5
Figura 3	Histograma do Sexo dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016	7
Figura 4	Histograma do Local de Residência dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016	8
Figura 5	Histograma da cor de Triagem atribuída às pulseiras dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016	8
Figura 6	Histograma dos diferentes destinos dos pacientes quando obtiveram alta do Hospital, entre Janeiro de 2012 e Dezembro de 2016	9
Figura 7	Gráfico do n ^o de admissões diárias em dias de feriado vs dias de não feriado, entre Janeiro de 2012 e Dezembro de 2016	11
Figura 8	recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016	12
Figura 9	Gráfico do número de admissões diárias no SUHB por Cor de pulseira atribuída na Triagem	13
Figura 10	Gráfico do número de admissões diárias no SUHB por Sexo do Paciente	14
Figura 11	Gráfico do número de admissões diárias no SUHB por Local de Residência do Paciente	15
Figura 12	Gráfico do número de admissões diárias no SUHB por destino do Paciente quando obteve alta do Serviço de Urgências	16
Figura 13	Gráfico do n ^o admissões diárias no SUHB, Janeiro 2012-Dezembro de 2016	41
Figura 14	Função de autocorrelação e de autocorrelação parcial dos resíduos do modelo	42
Figura 15	Função de autocorrelação e de autocorrelação parcial dos resíduos dos processo AR(7) e ARMA(2,1)	43
Figura 16	Gráfico dos Resíduos Padronizados ajustados pelo modelo AR(7) aos dados do SUHB	45
Figura 17	Gráfico Normal Quantil-Quantil e do Histograma dos resíduos obtidos pelo modelo AR(7)	45

Figura 18	Gráfico resíduos padronizados vs valores ajustados pelo modelo	46
Figura 19	Gráfico de comparação dos dados originais do SUHB vs dados previstos pelo modelo	48
Figura 20	Ilustração de previsão realizada pelo modelo 21	49
Figura 21	Gráfico de comparação dos dados originais do SUHB vs valores futuros para o mês de Dezembro 2016	51
Figura 22	Ilustração de previsão realizada pelo modelo 21	52
Figura 23	Gráfico de comparação dos dados originais do SUHB vs valores futuros (previstos diariamente) para o mês de Dezembro 2016	54
Figura 24	Widgets no package Shiny	56
Figura 25	Funções <i>render</i> no package Shiny	58
Figura 26	<i>Layout</i> da Web-App de Previsão do SUHB	60
Figura 27	Separador "Importação Base de Dados" da Web-App de Previsão do SUHB	61
Figura 28	Separador "Previsão Mensal" da Web-App de Previsão do SUHB	63
Figura 29	Separador "Previsão Diária" da Web-App de Previsão do SUHB	64
Figura 30	Separador "Estatísticas" da Web-App de Previsão do SUHB	65
Figura 31	Hospital Braga: Web-App Previsão	65

LISTA DE TABELAS

Tabela 1	Tabela dos agrupamentos definidos pela Administração Regional de Saúde	3
Tabela 2	Descrição variáveis da Base de Dados da Urgência Geral	6
Tabela 3	Tabela Resumo do nº de admissões diárias em dias de feriado vs dias não feriado, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	10
Tabela 4	Tabela Resumo do nº de admissões diárias em cada dia da semana, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	11
Tabela 5	Tabela Resumo do nº de admissões diárias em cada mês, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	12
Tabela 6	Tabela Resumo do nº de admissões diárias por cor de triagem, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	13
Tabela 7	Tabela Resumo de admissões diárias por sexo dos pacientes, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	13
Tabela 8	Tabela Resumo de admissões diárias por local de residência dos pacientes, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	13
Tabela 9	Tabela Resumo de admissões diárias por destino aquando alta hospitalar, entre 1 de Janeiro de 2012 até 31 Dezembro 2016	15
Tabela 10	Tabela identificação de processo ARMA comparando FAC vs FACP	34
Tabela 11	Tabela dos diferentes critérios de selecção dos modelos ARIMA sugeridos	44
Tabela 12	Tabela resumo do modelo do SUHB	47
Tabela 13	Tabela das estimativas dos coeficientes do processo AR(7)	48
Tabela 14	Tabela de Previsão de Y_t obtida do modelo (21) para o mês de Dezembro 2016	49
Tabela 15	Tabela de Previsão de Y_t obtida através da equação (22), para o mês de Dezembro de 2016	53

LISTA DE CÓDIGOS

6.1	Exemplo de documento UI Layout	57
6.2	Exemplo de documento Server	57
6.3	Exemplo de código a ser colocado no início do script R	59

INTRODUÇÃO

A realização deste trabalho foi possível graças à colaboração da Universidade do Minho com o Hospital de Braga. A equipa de gestão do Hospital está interessada em saber o número de admissões diárias no serviço de urgência geral de modo a tornar o work flow eficiente. Uma boa previsão das admissões no serviço permite ao Hospital gerir eficientemente toda a equipa que nele engloba, desde o número de médicos em serviço, o número de enfermeiros, etc.

Perante esta situação o objectivo deste trabalho é prever o número de pacientes que são admitidos no serviço de Urgência Geral do Hospital de Braga. Para isso teremos ao nosso dispôr uma base de dados, da qual, retiramos o número diário de admissões entre 1 de Janeiro de 2012 e 31 de Dezembro de 2016. A partir deste ponto criámos um modelo, com as seguintes variáveis explicativas: tempo, dia da semana, mês e dia de feriado ou não feriado.

Este trabalho contará com 5 Capítulos principais para além da Introdução e Conclusão.

O Capítulo 2, com o nome Serviço de Urgência do Hospital de Braga, apresenta uma pequena descrição do próprio Hospital. De seguida apresenta-se os vários serviços que o Hospital oferece, nomeadamente, o Serviço de Urgência Geral que é o foco deste trabalho. Na secção "Base de Dados" apresenta-se os dados da urgência geral, onde é realizada uma análise descritiva das várias variáveis presentes. Posteriormente, faz-se também uma análise do número de admissões diárias por dia de feriado, dia da semana, mês, cor de pulseira de triagem, sexo, local de residência e por destino depois de obtenção de alta do serviço acompanhado com os respectivos gráficos.

No Capítulo 3, com o nome Regressão Linear, apresenta-se os conceitos básicos de regressão linear. Apresenta-se como é feita a estimação por máxima verosimilhança para o modelo de regressão linear múltipla. Temos ainda os vários métodos/critérios de selecção de modelo, assim como, é realizado o diagnóstico de um modelo. Por fim temos as inferências num modelo de regressão.

No Capítulo 4, com o nome Modelos de Séries Temporais, faz-se uma breve descrição das diferentes metodologias de um modelo de séries temporais, também chamado de modelo ARIMA. Dividimos este Capítulo em três secções (Modelos estacionários lineares, Modelos não estacionários e Modelos de Séries temporais com covariáveis) que são essências para

a realização deste trabalho com especial atenção para a última secção. Nesta secção abordamos a metodologia de um modelo de regressão linear juntamente com um modelo de séries temporais que define o termo de erro desse modelo.

No Capítulo 5, com o nome Modelação dos Dados do Serviço de Urgência do Hospital de Braga, é o ponto central deste trabalho. Começa com a apresentação dos dados em estudo, de seguida é especificado o modelo de regressão. Temos duas subsecções onde estudamos duas equações de previsão. Os resultados obtidos da previsão vêm acompanhados pelos respectivos gráficos e tabelas de previsão. Posteriormente é feita uma comparação entre ambas as equações de previsão.

No Capítulo 6 é apresentada uma aplicação que foi desenvolvida com auxílio ao *software* R. Com a ajuda do *package* Shiny do R foi possível contruir a aplicação apresentada neste Capítulo, sendo que, a única função desta é tornar mais fácil ao utilizador aplicar os modelos apresentados neste trabalho. Para desta forma proceder à previsão.

Por último fazemos algumas observações e conclusões acerca do trabalho realizado.

SERVIÇO DE URGÊNCIA DO HOSPITAL DE BRAGA

O Hospital de Braga abriu portas em Maio de 2011, com uma capacidade de internamento até 705 camas. A construção deste Hospital substituiu o antigo Hospital de São Marcos.

Situa-se na Zona este da cidade de Braga. Tem uma arquitectura moderna de forma a permitir uma perfeita actividade dos diferentes serviços oferecidos pelo mesmo, por exemplo, os pacientes que são admitidos de urgência têm uma entrada própria não perturbando os serviços que estão em funcionamento no momento de admissão.

Um dos principais objectivos para a construção deste Hospital foi alargar os cuidados médicos para abranger mais pessoas dos distritos de Braga e Viana do Castelo.

O Hospital de Braga cobre uma área de cerca de 1.1 milhões de pacientes, provenientes de diferentes áreas do distrito de Braga e Viana do Castelo. A administração Regional de Saúde define os agrupamentos de centros de saúde (ACES) de acordo com as áreas geográficas. Este agrupamento está definido na Tabela 1.

Tabela 1: Tabela dos agrupamentos definidos pela Administração Regional de Saúde

ACES/Distrito	Local de Residência
ACES Cávado I	Braga
ACES Cávado II	Amares, Póvoa de Lanhoso, Terras de Bouro, Vieira do Minho e Vila Verde
ACES Cávado III	Barcelos e Esposende
ACES Alto Ave	Cabeceiras de Basto, Celorico de Basto, Fafe, Guimarães e Vizela
ACES Ave	Vila Nova de Famalicão
Viana do Castelo	Arcos de Valdevez, Caminha, Melgaço, Monção, Paredes de Coura, Ponte da Barca, Ponte de Lima, Valença, Viana do Castelo e Vila Nova de Cerveira
Outros	Outros concelhos/locais de residência de Portugal
Estrangeiro	Morada fora de Portugal

4 Capítulo 2. serviço de urgência do hospital de braga

O serviço de urgência do Hospital de Braga está dividido em três serviços:

1. Urgência Geral;
2. Urgência Obstétrica;
3. Urgência Pediátrica.

Neste trabalho iremos apenas analisar os dados referentes à urgência geral.

2.1 URGÊNCIA GERAL

O funcionamento do serviço de Urgências do Hospital de Braga é ilustrado pela Figura 1, ou seja, um paciente/utente que chegue às urgências do hospital segue os seguintes procedimentos:

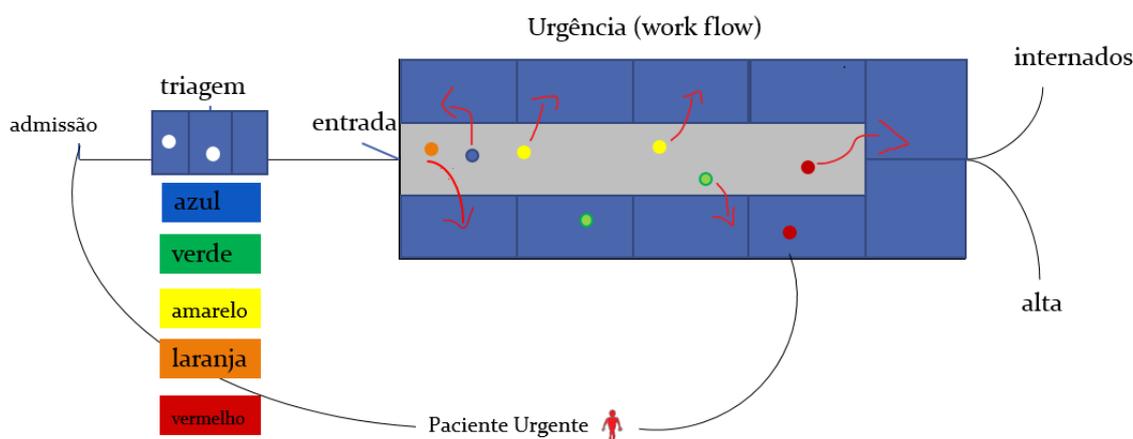


Figura 1: Diagrama Serviço de Urgências do Hospital de Braga

1. Admissão, é onde o paciente faz o seu registo de entrada;
2. Triagem, o paciente é triado de acordo com a triagem de manchester (Figura 2). Consiste na atribuição de uma cor (num total de cinco cores) com um certo tempo de espera segundo o tipo de urgência em questão.
3. Paciente Urgente/Emergente, é o tipo de paciente que se encontra num estado de emergência, é levado directamente para o serviço de tratamento do SUHB sem ser necessário fazer triagem.

4. Tratamento, alguns dos pacientes triados vão directamente para uma cama no serviço destinado para o seu tratamento, enquanto que, os restantes ficam em espera (dependendo da cor de triagem) até que uma nova cama esteja disponível. No entanto alguns pacientes deixam as urgências sem serem triados.
5. Internamento ou Alta, após o tratamento o médico de serviço indica ao paciente se este deve ser internado ou deve ter alta.

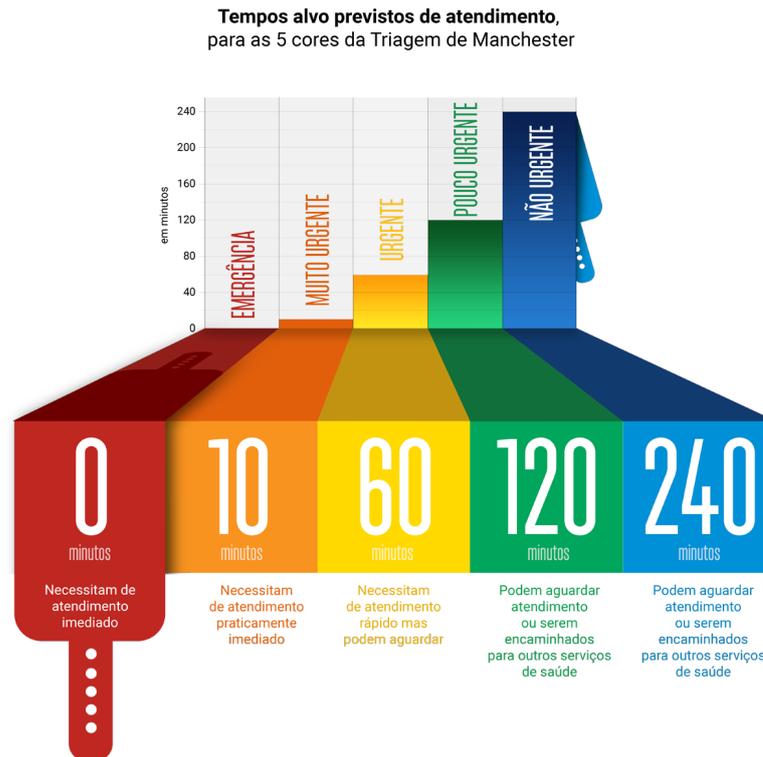


Figura 2: Tempo previsto até à 1^o observação médica segundo Triagem de Manchester (<http://www.grupoportuguestriagem.pt/>)

2.1.1 Base de Dados

A base de dados da Urgência Geral do Hospital de Braga contém informações de cerca de 650995 episódios. Desses episódios temos 16 variáveis que nos dão informações acerca da idade do paciente, data de entrada no serviço de urgência, etc. A tabela (2) dá-nos a descrição de cada uma das variáveis.

Como podemos observar pela tabela 2, temos que algumas das variáveis têm muitos níveis, de tal forma que, fez-se um agrupamento. A variável Concelho contém 271 níveis, e o agrupamento foi feito segundo o agrupamento definido pela administração Regional de Saúde (tabela (1)).

Tabela 2: Descrição variáveis da Base de Dados da Urgência Geral

Descrição da variável	Tipo Variável/Valores	Nome
Dá-nos informações referentes a cada paciente	Quantitativa	Episódio
Indica o local para o qual o paciente foi admitido	Qualitativa (1 nível)	Local
Indica o sexo do paciente	Qualitativa (2 níveis)	Sexo
Indica a idade do paciente	Quantitativa	Idade
Indica data de nascimento do paciente	Qualitativa (data)	Data.Nascimento
Indica o concelho de residência do paciente	Qualitativa (294 níveis)	Concelho
Indica data de admissão do paciente	Qualitativa (data)	Data.Admissão
Indica a hora de admissão do paciente	Qualitativa (hora)	Hora.Admissão
Indica data de alta do paciente	Qualitativa (data)	Data.Aalta
Indica hora de alta do paciente	Qualitativa (hora)	Hora.Aalta
Indica cor de pulseira atribuída ao paciente	Qualitativa (6 níveis)	Cor.Triagem
Indica data de triagem do paciente	Qualitativa (data)	Data.Triagem
Indica hora de triagem do paciente	Qualitativa (hora)	Hora.Triagem
Indica data da 1ª observação do paciente	Qualitativa (data)	Data.1ª.Obs
Indica hora da 1ª observação do paciente	Qualitativa (hora)	Hora.1ª.Obs
Indica o destino do paciente quando obteve alta	Qualitativa (15 níveis)	Destino

Outra variável que sofreu alterações foi a variável Destino que contém 15 níveis, passa a ter 5 níveis ("Abandono", "Exterior", "Falecido", "Outro Hospital" e "Serviço de Internamento").

Já a variável Cor de Triagem contém 5 níveis, referentes à cor das pulseiras atribuídas aos pacientes aquando a sua admissão, que no entanto, passará a conter 6 níveis. Este sexto nível será referente aos pacientes que são admitidos no serviço de urgência que acabam por abandonar o Hospital.

Algumas das variáveis presentes na tabela anterior não irão ser de interesse para o nosso trabalho. Elas são as datas e horas, com exceção da variável "Data.Admissão". Esta será importante para conseguirmos retirar o número de admissões diárias no SUHB, cujo estudo será realizado no Capítulo 5.

Os dados do SUHB são referentes ao período de 1 de Janeiro de 2012 até 31 de Dezembro de 2016, no qual foram admitidos 352991 (54.2%) pacientes do sexo feminino e 298004 (45.8%) do sexo masculino - ver Figura 3.

Pela Figura 4 observamos que a maioria dos pacientes admitidos no SUHB são provenientes do agrupamento ACES Cávado I, ou seja, de Braga (como era de esperar!). Desse agrupamento foram admitidos 324589 (49.9%) pacientes, quase metade do total dos pacientes dos restantes agrupamentos.

No entanto, o agrupamento ACES Cávado II foi o segundo com mais admissões (com cerca de 193507 (29.7%) admissões), que é referente aos concelhos de Vila Verde, Póvoa de Lanhoso, Amares, Terras de Bouro, tudo concelhos circundantes da cidade de Braga. Os restantes agrupamentos ACES Cávado III, Estrangeiro, Viana do Castelo, ACES Alto Ave, ACES Ave e Outros obtiveram, respectivamente, 55680 (8.6%), 892 (0.1%), 15689 (2.4%), 25979 (4%), 19332 (3%) e 15327 (2.4%).

A Figura 5, representa o histograma do número de admissões por cor de pulseira de triagem. A cor de pulseira mais atribuída aos pacientes que recorreram ao SUHB foi a cor Amarela com cerca de 390671 (60%) admissões, em segundo lugar temos a cor Verde com 191257 (29.4%) admissões. As restantes cores foram as menos atribuídas, Azul (0.8%), Laranja (8.9%), Vermelho (0.5%). Por último temos aqueles pacientes que abandonam o SUHB sem serem triados, esta situação ocorreu 2221 (0.3%) vezes.

Em relação ao destino dos pacientes que obtêm alta do Hospital, temos que, cerca de 551315 (84.7%) são reencaminhados para outros serviços do Hospital como consultas de especialidade, para centros de saúde, etc. Do total de pacientes que foram admitidos apenas 61787 (9.5%) foram internados, 15432 (2.4%) foram enviados para outros hospitais e 1666 (0.3%) faleceram.

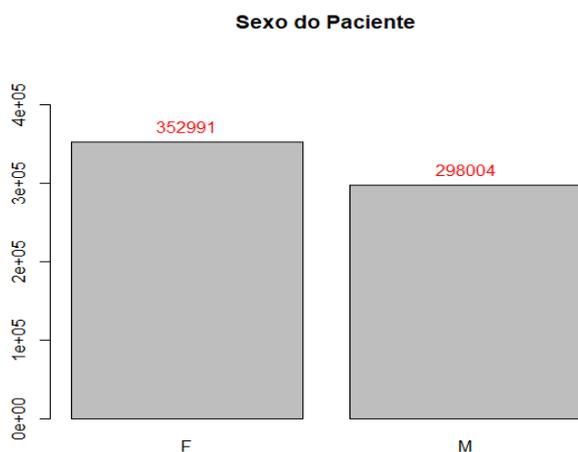


Figura 3: Histograma do Sexo dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016

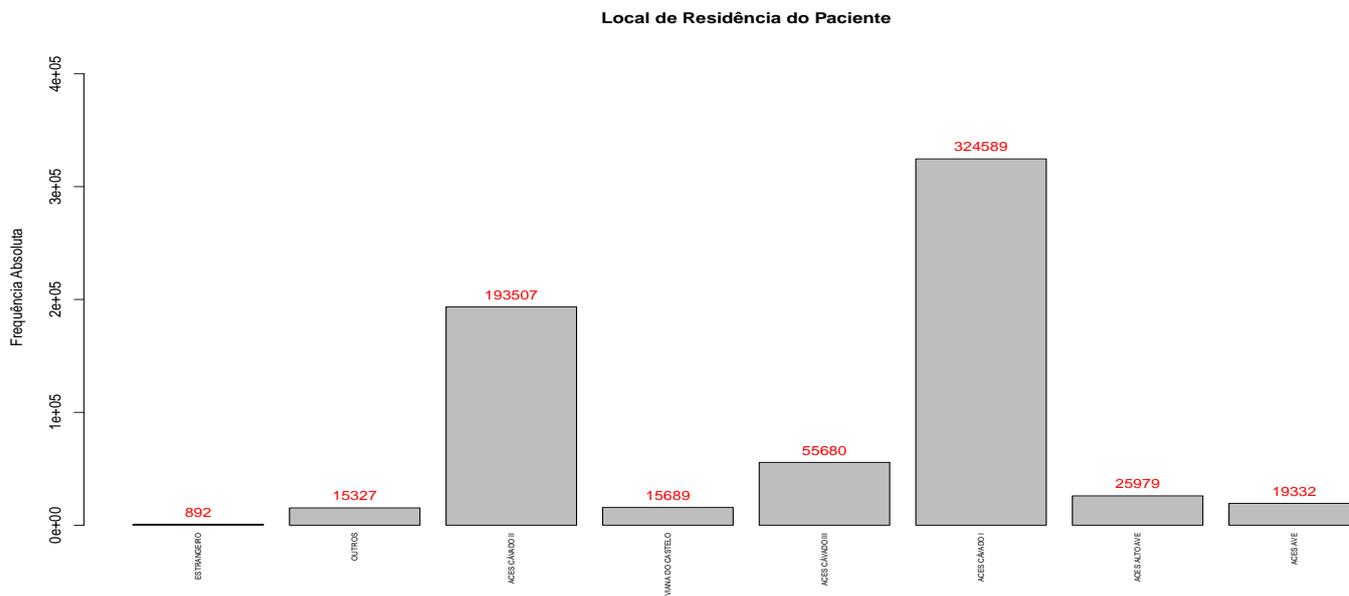


Figura 4: Histograma do Local de Residência dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016

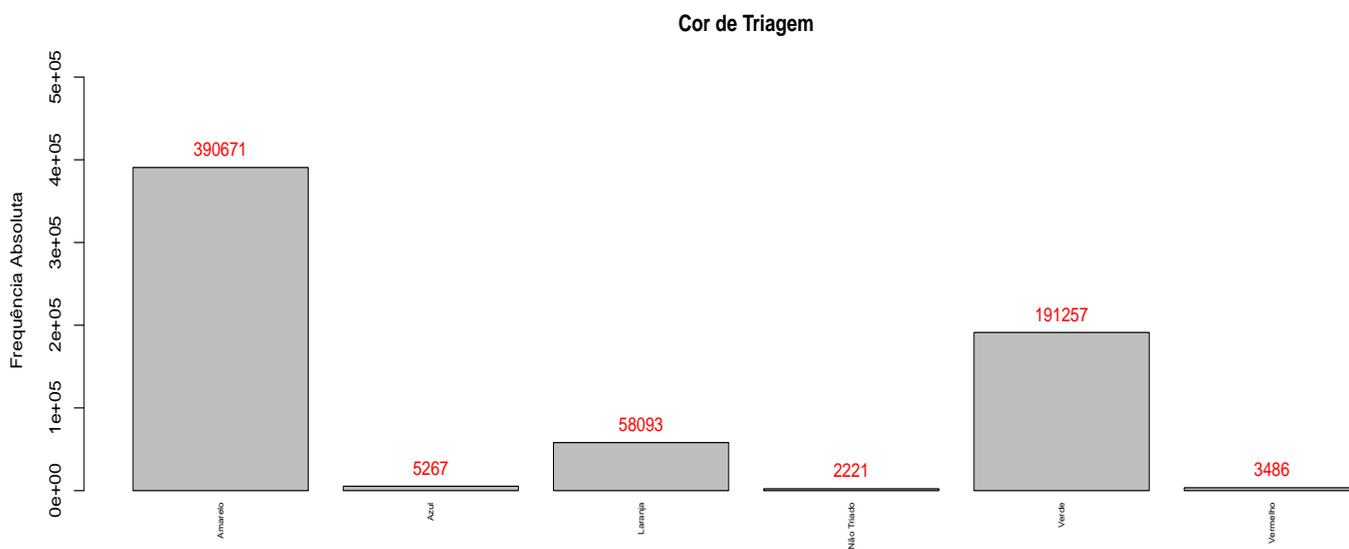


Figura 5: Histograma da cor de Triagem atribuída às pulseiras dos pacientes que recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016

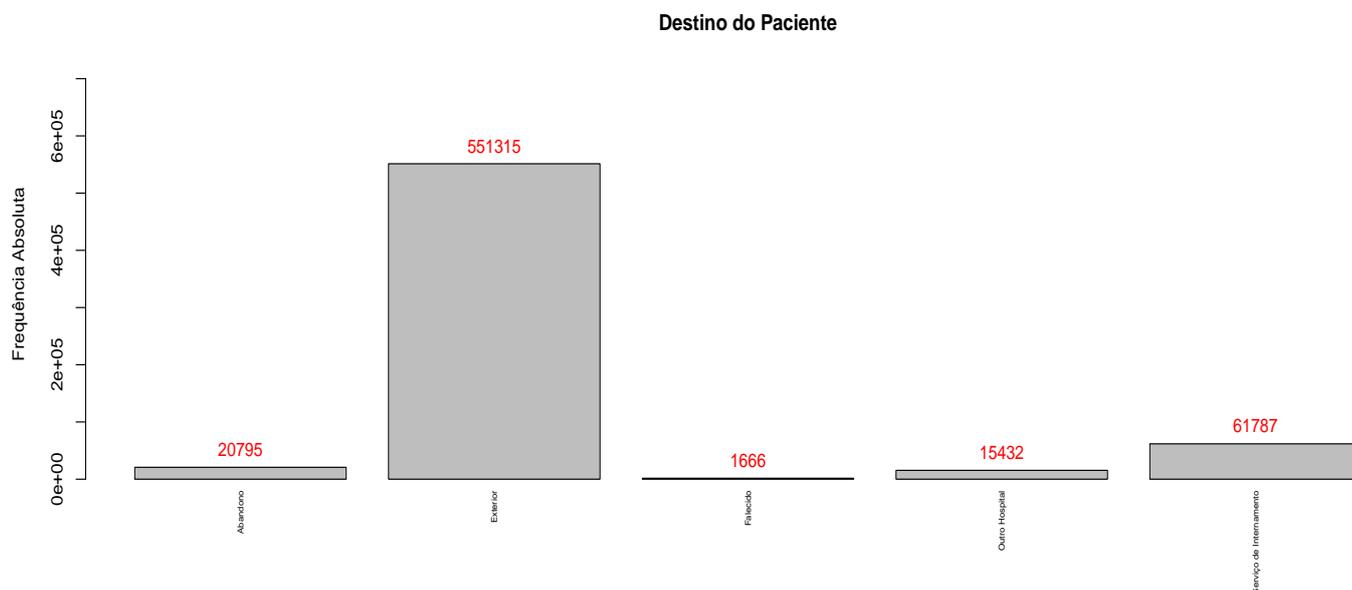


Figura 6: Histograma dos diferentes destinos dos pacientes quando obtiveram alta do Hospital, entre Janeiro de 2012 e Dezembro de 2016

Nas tabelas (3), (4), (5), (6), (7), (8) e (9) faz-se a descrição do número de admissões diárias por dia da semana, mês, por dia de feriado ou não, cor de triagem, local de residência e destino do paciente quando obteve alta do SUHB.

O número de admissões diárias em dias de feriado é menor que em dias em que não há feriados, ou seja, são admitidos em média 285 pacientes contra os 327 pacientes dos dias em que não há feriado. O gráfico 7 corrobora isso mesmo.

A tabela 4 referente ao número de admissões diárias em cada dia da semana diz que o número de pacientes que recorre ao SUHB diminui ao longo da semana, começando na segunda-feira e terminando no Domingo. Temos que nas segundas-feiras o número de pacientes é maior - em média temos cerca de 407 pacientes - e no domingo temos em média cerca de 299 pacientes. Pelo gráfico da Figura 8 confirmamos que segunda-feira é o dia com mais admissões diárias, já de terça-feira a sexta-feira o número de admissões parece manter-se no mesmo nível. No sábado o número médio de admissões é de 326 pacientes com um máximo de 432 pacientes. No Domingo temos em média cerca de 299 pacientes num mínimo de 204 e um máximo de 387 pacientes.

A tabela (5) mostra o número de admissões diárias por cada mês do ano. A primeira observação a fazer é que o número de admissões médias permanece constante ao longo dos meses, com cerca de 350 admissões médias mensais, o mesmo acontece com o número total de admissões por cada mês com cerca de 55 mil admissões. O mês de Dezembro tem o maior número de admissões diárias (total de 55169 pacientes) com uma média de 355 pacientes e com o seu valor máximo de cerca 510 e mínimo de 237.

A tabela 6 representa o número de admissões diárias por cor de pulseira atribuída aos pacientes no momento da triagem. Temos que a cor de pulseira Amarelo, Verde e Laranja são as mais utilizadas com 390671 admissões para a cor Amarelo, 191257 para a cor Verde e 58093 para a cor Laranja. As cores Azul (5267 admissões) e Vermelho (3486 admissões) foram as que menos foram utilizadas sendo que diariamente a cor azul foi atribuída em média a cerca de 3 pacientes enquanto que a cor vermelho foi atribuída a cerca de 2 pacientes. Já as cores de pulseira Amarelo, Verde e Laranja foram atribuídas diariamente a cerca de 214, 195 e 32 pacientes, respectivamente. A figura 9 representa os dados da tabela.

A tabela 7 representa o número de admissões diárias por sexo do paciente. Podemos observar que o sexo Feminino tem um número de admissões diárias superior ao sexo Masculino, sendo que cada sexo tem respectivamente um número médio de admissões de 193 e 163 pacientes num total de 352991 e 298004 admissões entre Janeiro de 2012 e Dezembro de 2016, ver Figura (10).

A tabela 8 apresenta o número de admissões diárias por local de residência do paciente. Os locais de residência pertencentes aos agrupamentos ACES Cávado I, ACES Cávado II e ACES Cávado III têm o maior número de admissões diárias, uma vez que, esses agrupamentos contém os concelhos mais próximos do hospital de Braga. A figura 11 representa os dados da tabela.

Por último temos a tabela 9 representa os vários destinos dos pacientes quando estes obtêm alta do serviço de urgência. Temos que a maioria dos pacientes admitidos diariamente são reencaminhados para outros serviços do hospital, em média cerca de 302 pacientes. Cerca de 34 pacientes que entram no SUHB são internados no hospital, 12 pacientes acabam por abandonar o serviço e 9 pacientes são enviados para outro hospital. A figura 12 representa os dados da tabela 9.

Tabela 3: Tabela Resumo do nº de admissões diárias em dias de feriado vs dias não feriado, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Feriado	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Sim	204	285	314	311.5	336	406	17753
Não	221	327	358	357.8	388	510	633242

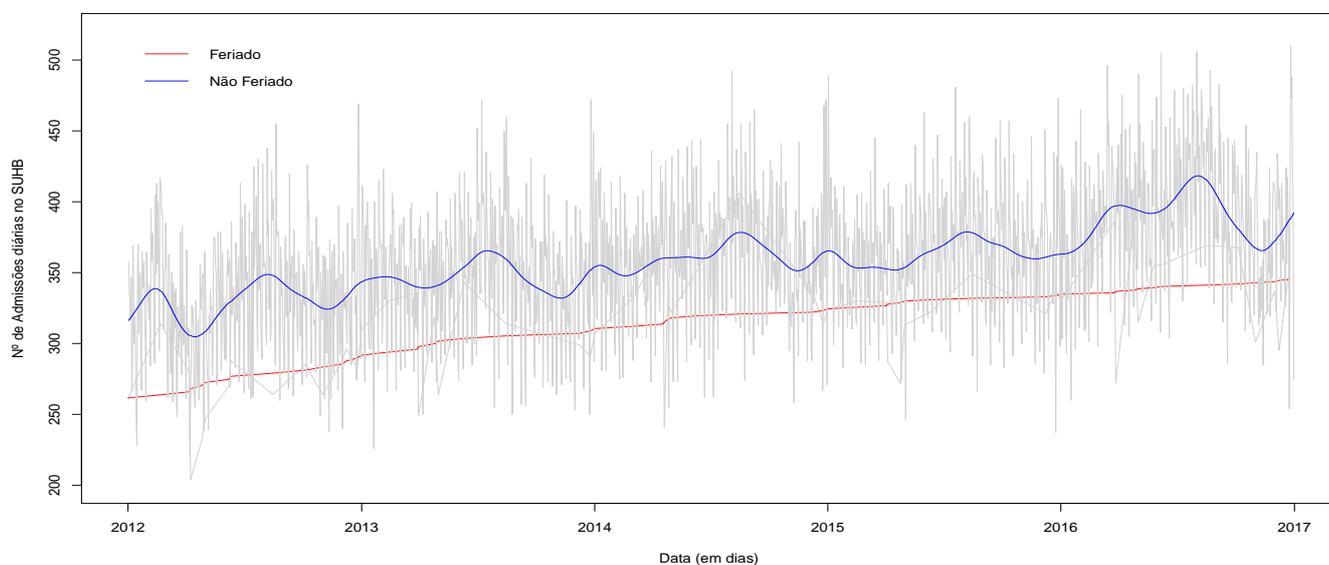


Figura 7: Gráfico do nº de admissões diárias em dias de feriado vs dias de não feriado, entre Janeiro de 2012 e Dezembro de 2016

Tabela 4: Tabela Resumo do nº de admissões diárias em cada dia da semana, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Dia da Semana	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Segunda-Feira	274	379	406	407.3	434	510	106294
Terça-Feira	247	342	369	367.6	393	473	95947
Quarta-Feira	231	342	368	365.8	388	488	95464
Quinta-Feira	237	342	363	363.2	387	472	94786
Sexta-Feira	276	342	362	364.7	383	489	95183
Sábado	221	303	325	326.5	349	432	85220
Domingo	204	279	297	299.2	320	387	78101

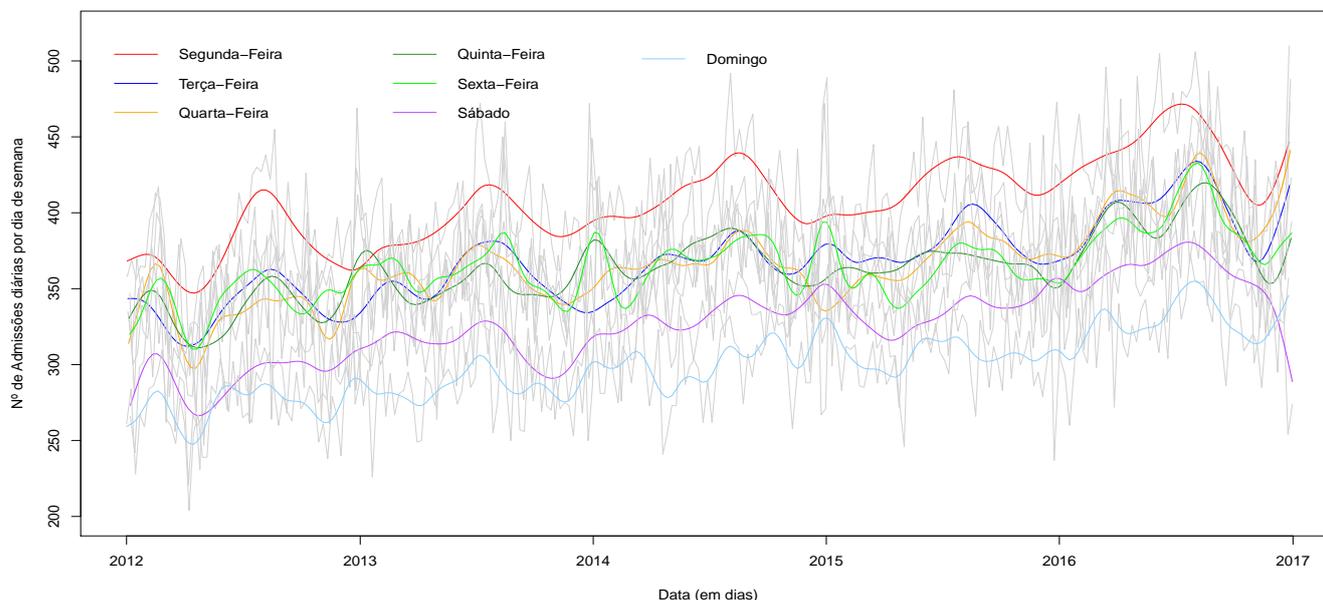


Figura 8: recorreram ao SUHB, entre Janeiro de 2012 e Dezembro de 2016

Tabela 5: Tabela Resumo do nº de admissões diárias em cada mês, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Mês	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Janeiro	226	318	354	350	384.5	489	54246
Fevereiro	266	328.2	352.5	352.6	377.8	465	50066
Março	248	313	354	350.6	382.5	496	54346
Abril	204	316.2	350.5	345.6	382	475	51833
Maio	239	324.5	352	353.3	379.5	490	54758
Junho	262	324	360.5	359.2	388.8	505	53873
Julho	255	338.5	374	371.9	402.5	482	57642
Agosto	250	350	374	377.8	417	506	58558
Setembro	256	325.2	362	360	388	483	54004
Outubro	249	332	355	356.9	381.5	457	55314
Novembro	238	315	344	341.2	370	446	51186
Dezembro	237	323	355	355.9	392	510	55169

Tabela 6: Tabela Resumo do nº de admissões diárias por cor de triagem, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Cor de Triagem	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Amarelo	108	194	213	213.8	233	308	390671
Azul	0	1	2	2.883	4	17	5267
Verde	34	88	104	104.7	120	202	191257
Laranja	11	27	31	31.8	36	56	58093
Vermelho	0	1	2	1.9	3	8	3486
Não Triado	0	0	1	1.216	2	24	2221

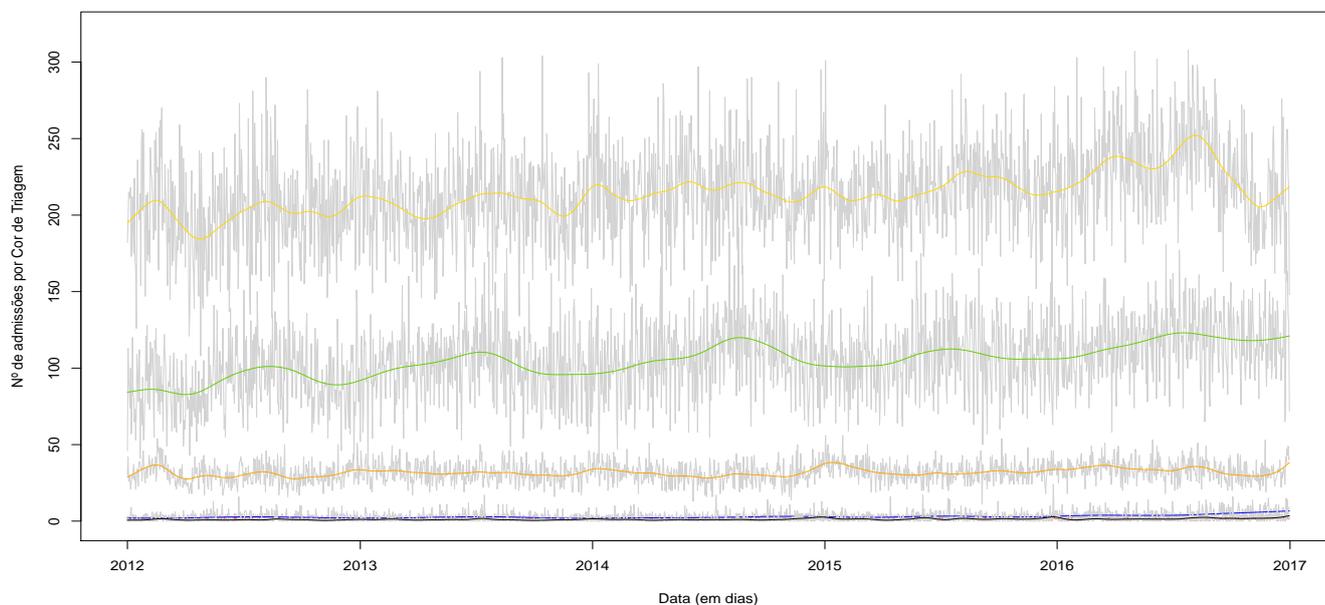


Figura 9: Gráfico do número de admissões diárias no SUHB por Cor de pulseira atribuída na Triage

Tabela 7: Tabela Resumo de admissões diárias por sexo dos pacientes, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Sexo	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Masculino	93	148	163	163.1	177	245	298004
Feminino	105	172	194	193.2	214	289	352991

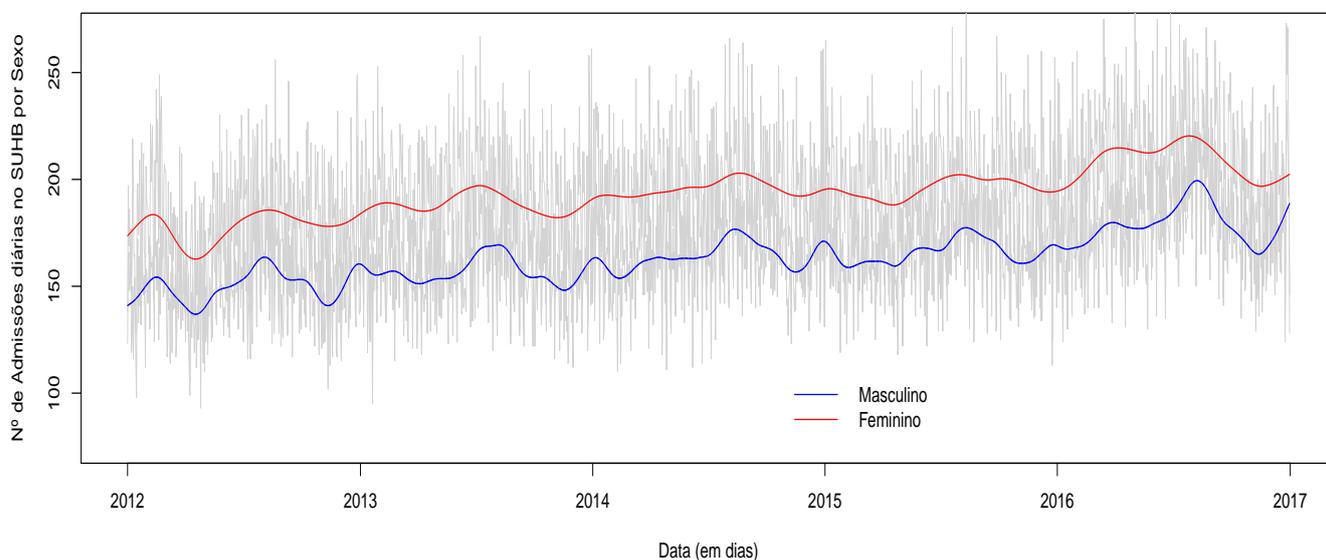


Figura 10: Gráfico do número de admissões diárias no SUHB por Sexo do Paciente

Tabela 8: Tabela Resumo de admissões diárias por local de residência dos pacientes, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Local de Residência	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
ACES Cávado I	106	160	177	177.7	194	259	324589
ACES Cávado II	54	94	106	105.9	117	164	193507
ACES Cávado III	9	25	30	30.48	35	62	55680
Viana de Castelo	0	6	8	8.587	11	22	15689
Estrangeiro	0	0	0	0.4882	1	9	892
Outro	0	6	8	8.389	10	29	15327
ACES Alto Ave	4	11	14	14.22	17	34	25979
ACES Ave	1	8	10	10.58	13	25	19332

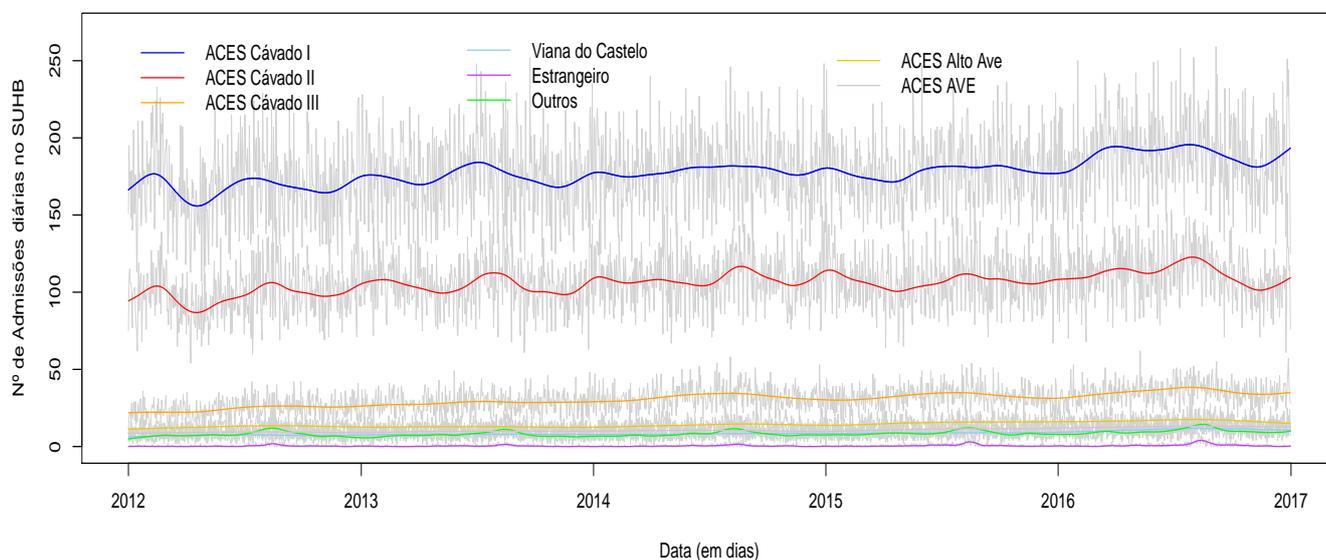


Figura 11: Gráfico do número de admissões diárias no SUHB por Local de Residência do Paciente

Tabela 9: Tabela Resumo de admissões diárias por destino aquando alta hospitalar, entre 1 de Janeiro de 2012 até 31 Dezembro 2016

Destino	Min	1º Qua	Mediana	Média	3º Qua	Máx	Total
Serviço de Internamento	15	29	34	33.82	39	63	61787
Abandono	0	6	9	11.38	15	68	20795
Exterior	164	276	302	301.8	328	427	551315
Falecido	0	0	1	0.91	1	6	1666
Outro Hospital	0	6	8	8.447	10	20	15432

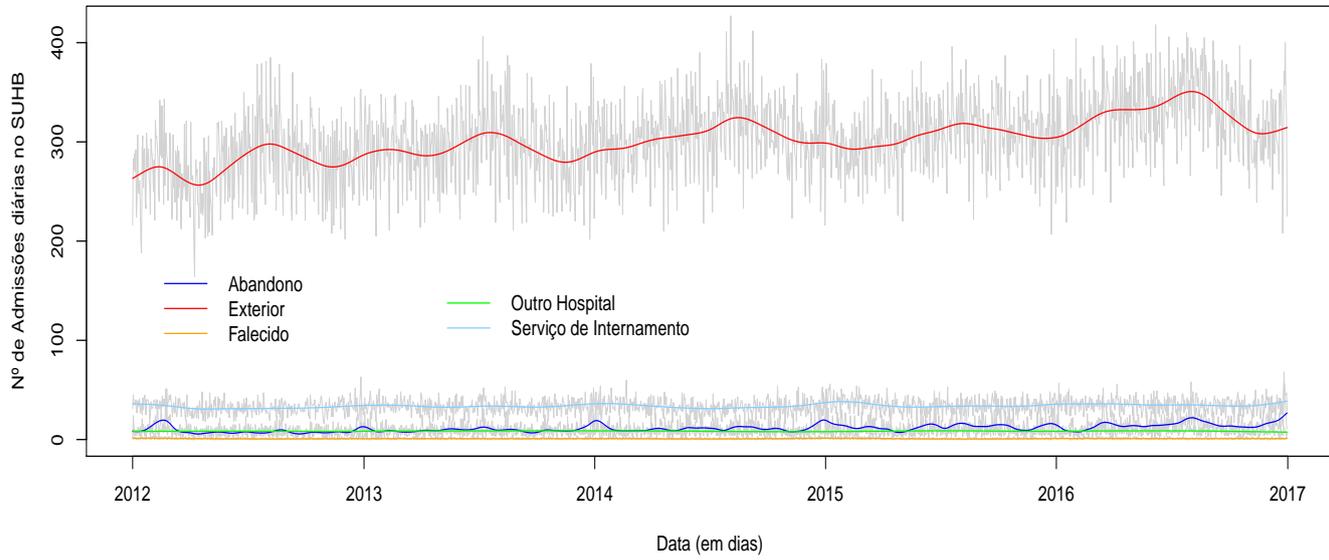


Figura 12: Gráfico do número de admissões diárias no SUHB por destino do Paciente quando obteve alta do Serviço de Urgências

REGRESSÃO LINEAR

Neste capítulo será abordado os conceitos básicos de regressão linear, assim como regressão múltipla.

A palavra *regressão* significa "*regressar a um lugar ou estado anterior*". Portanto quando utilizamos a metodologia de regressão linear, queremos estimar o melhor modelo possível baseado nos dados observados.

Na análise de regressão é assumido que a população (Y) pode estar linearmente relacionada com variáveis de entrada (X 's) por coeficientes fixos (β 's).

3.1 NOTAÇÃO

Nesta seção definimos a notação que irá ser utilizada ao longo deste capítulo. Para tal segue-se o seguinte:

- Y_i - Refere-se à nossa variável resposta, que no caso deste trabalho irá ser referente ao número de admissões/pacientes diária(o)s no SUHB. Que irá depender de uma ou mais variáveis ou eventos.
- $X_{1,i}, \dots, X_{p,i}$ - Refere-se à variável ou variáveis que podem afetar Y . Temos $X_{p,i}$ variáveis em que p é o número de variáveis do modelo, $i = 1, \dots, n$ e $n \in \mathbb{N}$. Neste trabalho iremos ver que as variáveis $X_{p,i}$ representam respetivamente: uma variável binária que indica se é um dia de feriado ou não, uma variável com sete níveis que indica o dia da semana (segunda-feira,...) e uma variável com doze níveis que indica o mês do ano (Janeiro,...).
- i - É utilizado para indicar a i -ésima observação: Y_i e $X_{p,i}$ refere-se a variáveis cujos valores estão ordenados no tempo. Temos que $i = 1, 2, 3, \dots, n$, onde i indica a i -ésima observação correspondente na sequência dos nossos dados. Ou seja, Y_1 corresponde

à primeira observação da variável Y , e podemos continuar assim até ao último valor de i .

3.2 MODELOS DE REGRESSÃO LINEAR

Segundo [Dobson \(2002\)](#) para formular um modelo temos de ter em conta duas componentes:

- Distribuição de Probabilidade de Y , que neste trabalho será $Y \sim N(\mu, \sigma^2)$;
- e a equação que liga o valor esperado de Y com a combinação linear das variáveis explicativas: $E[Y] = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots$

No caso de regressão linear simples, o modelo é dado por:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad i = 1, \dots, N \quad (1)$$

Chamamos à equação (1) função de regressão populacional, ou seja, assumimos que a variável resposta populacional (Y) pode estar linearmente relacionada com uma ou mais variáveis explicativas ($X_{1,i}$) através de coeficientes fixos. β_0 e β_1 são os parâmetros fixos a serem estimados, ϵ representa o termo de erro e está presente no modelo porque a relação entre $X_{1,i}$ e Y_i não é total. β_1 é a média populacional que é alterada em Y_i à medida que aumentamos uma unidade em $X_{1,i}$ e N representa o tamanho da população.

Quando temos mais do que uma variável explicativa estamos perante um modelo de regressão múltipla, cujo modelo é dado por,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_p X_{p,i} + \epsilon_i \quad i = 1, 2, \dots, N \quad (2)$$

onde p corresponde ao número de variáveis explicativas, i corresponde à posição da observação da variável explicativa (ex: $X_{2,8}$ é o valor da oitava observação da segunda variável explicativa), N corresponde ao tamanho da população - n quando estamos perante uma amostra. Também é possível escrever a equação (2) na forma matricial

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e} \quad (3)$$

onde \mathbf{Y} é um vector $n \times 1$ com n observações da amostra, \mathbf{X} é uma matriz $n \times k$ de n observações em $k - 1$ variáveis, mais uma coluna com 1's que representa β_0 , \mathbf{B} é um vector $k \times 1$ com os coeficientes do modelo e \mathbf{e} é um vector $n \times 1$ com os resíduos (onde $e_i = Y_i - \hat{Y}_i$ ou na forma matricial $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$).

3.2.1 Estimação

A estimação dos coeficientes de regressão para o modelo de regressão linear simples será realizada através do método de mínimos quadrados, ao qual iremos denotar por Q , sendo que este método de estimação é um caso particular do método de estimação por máxima verosimilhança. Utilizamos este método porque queremos minimizar a soma do erro ao quadrado $\sum \epsilon^2$ - se representarmos a reta graficamente, corresponde à linha vertical entre o valor observado Y_i e o valor estimado \hat{Y}_i . Para tal segue a seguinte equação:

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{1,i}))^2 \quad (4)$$

Através de (4) conseguimos obter a equação estimada de cada um dos coeficientes do modelo, derivando a equação em ordem a cada um dos β 's. Obtemos as seguintes equações,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_{1,i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{1,i} - \bar{X}_1)^2} = \frac{Cov(X_1, Y)}{S_{X_1}^2} \quad (5)$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}_1 \quad (6)$$

onde \bar{Y} e \bar{X} representa a média de Y_i e $X_{1,i}$. n diz respeito ao número de observações disponíveis e o indicador i representa a localização da observação numa variável (ex: $X_{1,6}$ representa a sexta observação da variável X_1) e $S_{X_1}^2 = \sum_{i=1}^n (X_{1,i} - \bar{X}_1)^2$.

Propriedades Estimadores:

- | | |
|---|--|
| 1. $E[\hat{\beta}_1] = \beta_1$ | 4. $E[\epsilon] = 0$ |
| 2. $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_{1,i} - \bar{X}_1)^2}$ | 5. $Var(\epsilon) = \sigma_\epsilon^2$ |
| 3. $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{X}_1}{\sum_{i=1}^n (X_{1,i} - \bar{X}_1)^2}$ | 6. $E[\hat{\sigma}^2] = \sigma^2$ |

Em regressão linear múltipla, a estimação dos parâmetros do modelo é feita por máxima verosimilhança. Vejamos que a função de densidade de probabilidade de Y_i é dada por:

$$f_i(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \beta_1 X_{1,i} - \dots - \beta_p X_{p,i})^2\right)$$

daqui, concluímos que a função de verosimilhança é

$$\begin{aligned} L(\mathbf{B}|\mathbf{Y}) &= \prod_{i=1}^n f_i(Y_i) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_1 X_{1,i} - \dots - \beta_p X_{p,i})^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{1}{2\sigma^2} |\mathbf{Y} - \mathbf{XB}|^2 \right) \end{aligned}$$

Antes de maximizar a função de verosimilhança, temos de primeiro, minimizar $|\mathbf{Y} - \mathbf{XB}|^2$. Só depois podemos derivar a função verosimilhança em ordem a cada β igualando a zero.

O vector dos coeficientes, na forma matricial, estimados a partir do método de máxima verosimilhança é dado por

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

com matriz de variância-covariância dada por,

$$\text{Var} - \text{Cov}(\hat{\mathbf{B}}) = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

3.2.2 Seleção de Modelo

Quando se trabalha com um modelo de regressão linear múltipla, está-se a lidar com várias variáveis explicativas e de tal forma pretende-se obter o melhor modelo que se ajusta aos dados. Um dos métodos utilizados é adicionar e remover variáveis explicativas, a este procedimento chamamos método de selecção *stepwise*. Este método é baseado no critério de informação de Akaike que começa por adicionar variáveis ao modelo, depois calcula o seu valor de AIC e introduz ou remove essa variável caso esse valor for o menor possível que inicialmente tinha com o modelo nulo (refere-se ao modelo que não contém variáveis explicativas).

Critério de Informação de Akaike (AIC)

A utilização da estatística AIC é das mais utilizadas para fazer a escolha de selecção de um modelo. Tem em conta a equação 3, o AIC é dado por

$$AIC = -2\log(\hat{L}) + \alpha q$$

onde q representa o número de parâmetros β desconhecidos no modelo, L representa a verosimilhança estimada e α é uma constante pré-determinada (toma um valor entre 2 e 6,

se o valor escolhido for 3 corresponde a escolher um nível de significância a 5%). Quanto menor for o valor de AIC , melhor será o ajustamento do modelo aos dados.

Coefficiente de Determinação (R^2)

Por vezes os dados nem sempre apresentam uma relação linear, de tal forma que, deveríamos ter uma medida que expressa-se a força da relação linear entre as variáveis explicativas e a variável resposta. Para tal definimos a seguinte equação:

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

R^2 representa a percentagem de variabilidade em Y - chamamos de coeficiente de determinação. Se tivermos $\hat{\beta}_1 = 0$ então $\hat{Y} = \bar{Y}$ e $R^2 = 0$, por outro lado se $Y = \hat{Y}$ então $R^2 = 1$. R^2 varia entre 0 e 1, no entanto, queremos que R^2 seja o mais próximo possível de 1.

Esta medida é alvo de muitas críticas porque à medida que aumentamos o número de variáveis o seu valor aumenta sendo estas variáveis significativas ou não. Consequentemente utilizamos o coeficiente de determinação ajustado para melhor explicar a variabilidade de Y , ao qual denotamos por R_a^2 .

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

onde $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, $SST = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$, n é o número de observações e p o número de variáveis explicativas.

Factor de Inflação da Variância (VIF)

Quando temos certas variáveis explicativas altamente correlacionadas chamamos de colinearidade ou multicolineariedade. Quando isto acontece a matrix de desenho \mathbf{X} pode estar linearmente dependente, ou seja, significa que se forem feitas pequenas alterações nos dados a estimação de $\hat{\mathbf{B}}$ será altamente influenciada. Por consequência os valores da matrix de variância-covariância serão grandes dando variâncias e covariâncias altas para os valores dos β 's. Colinearidade significa que escolher as variáveis explicativas que melhor explicam a resposta é difícil.

Para tal calculamos o factor de inflação da variância, o qual denotamos por VIF , para cada variável explicativa.

$$VIF_j = \frac{1}{1 - R_j^2}$$

onde R_j^2 é o coeficiente de determinação obtido ao fazer a regressão das j -ésimas variáveis explicativas. Se não estão correlacionadas então o $VIF = 1$, ou seja se $VIF = 10$ ou superior teremos motivos para estar preocupados.

O VIF aumenta à medida que aumenta a correlação entre as variáveis. Se muitas das variáveis explicativas estão correlacionadas então não é possível saber qual deve ser incluída no modelo. Num caso destes teremos de estudar bem os dados e perceber qual as variáveis que mais é interessante colocar no modelo. (Dobson, 2002)

3.2.3 MAPE

O erro percentual absoluto médio (MAPE) é uma medida relativa, e é dado por

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|e_t|}{Y_t} \times 100$$

onde n é o tamanho da amostra, e_t é o erro previsto pelo modelo no tempo t . Lewis (1982) estabeleceu um critério de interpretação desse valor, ou seja, no caso em que o valor de MAPE é inferior a 10% a previsão é altamente precisa; entre 10-20% a previsão é boa; 20-50% a previsão é razoável e maior que 50% a previsão é má.

3.2.4 Diagnóstico

Nesta secção irá ser discutido a verificação do modelo, ou seja, verifica se o modelo se ajusta bem aos dados. Para tal verifique-se os seguintes pressupostos do modelo:

- Normalidade dos resíduos, $\epsilon_i \sim N(0, \sigma_\epsilon^2)$;
- Independência, ϵ_i 's são independentes;
- Variância constante, $Var(\epsilon_i) = \sigma_\epsilon^2$;
- Média Zero, $E[\epsilon_i] = 0$.

Os resíduos do modelo representado na equação (2) e (3) são definidos por

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}^T \hat{\mathbf{B}} \quad (7)$$

Segundo Ryan (1997) é preferível estudar os resíduos padronizados do que os resíduos, $\hat{\epsilon}_i$, representados na equação (7). No entanto, os resíduos padronizados têm uma vantagem

sobre os resíduos: estão reduzidos a uma escala de forma a serem independentes de Y e da sua unidade de medida e são dados pela seguinte expressão:

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}(1 - h_{ii})^{\frac{1}{2}}}$$

onde h_{ii} é o i -ésimo elemento da diagonal da matriz de projecção ou em inglês *hat matrix*, que por sua vez é dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

Verificando o primeiro pressuposto, ou seja, normalidade dos resíduos é representado o gráfico dos resíduos contra os valores esperados. Outra forma de verificar é realizar o teste de Shapiro Wilks.

Representando os resíduos padronizados (r_i) contra os valores ajustados pelo modelo (\hat{Y}_i) é possível detectar alterações na variância. Se ocorrer um aumento na forma como os resíduos estão espalhados à medida que se aproximam do final do limite dos valores ajustados, pode indicar que a variância não é constante - homocedasticidade.

Para verificar a independência e média, representa-se o gráfico dos resíduos sendo que estes devem estar apresentados de forma aleatória e sem qualquer padrão aparente à volta da média zero. (Dobson, 2002)

Outras medidas de diagnóstico

Ainda segundo Dobson (2002), existem outros métodos de verificar o modelo:

1. Pontos Outliers: é uma observação que não é bem ajustada pelo modelo, tem valor mais elevado que as restantes observações.
2. Pontos Influentes: são observações que têm um grande efeito nas inferências baseadas no modelo.
3. Pontos Alavanca: o valor h_{ii} , i -ésimo elemento da diagonal da matriz de projecção. Se h_{ii} for maior que dois ou três o valor de p/N é motivo para preocupação (p é o número de parâmetros e N é o número de observações)
4. Resíduos padronizados e pontos alavanca dão origem a outras estatísticas, como a distância de Cook

$$D_i = \frac{1}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right) r_i^2$$

3.2.5 Inferências no Modelo de Regressão

Teste de Hipóteses para β_p

De forma a podermos saber se um parâmetro do modelo é ou não significativo, recorreremos a um teste de hipóteses. Tanto para regressão linear simples e para regressão linear múltipla esse teste é dado pela seguinte expressão,

$$\begin{aligned} H_0 : \beta_p &= b_p \\ &vs \\ H_1 : \beta_p &\neq b_p \end{aligned}$$

com estatística de teste dada por

$$ET = \frac{\hat{\beta}_p - b_p}{\sqrt{\hat{\sigma}^2 C_{pp}}} = \frac{\hat{\beta}_p - b_p}{\sqrt{Var(\hat{\beta}_p)}} \sim t_{n-p-1}$$

onde C_{pp} é o p -ésimo elemento da diagonal da matriz $C = (\mathbf{X}^T \mathbf{X})^{-1}$ e p corresponde ao número de variáveis explicativas do modelo.

Daqui tiramos o valor do p -valor de forma a rejeitar ou aceitar a hipótese nula. O p -valor é dado por $2 \times (1 - P(t_{n-p-1} > |ET|))$ ou temos que a região de rejeição é dada por $RC = \{ET : |ET| > t_{\frac{\alpha}{2}, n-p-1}\}$ em que α é o nível de significância.

Regressão Linear Simples*Intervalo de Confiança para β_1*

O intervalo de confiança para β_1 a $(1 - \alpha)\%$ de confiança, com distribuição t -Student com $n - 2$ graus de liberdade é dado por,

$$\left[\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{Var(\hat{\beta}_1)} \right]$$

ou seja, o limite inferior é dado por $\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$ e o limite superior dado por $\hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$, α é o nível de significância.

Intervalo de Confiança para β_0

O intervalo de confiança para β_0 a $(1 - \alpha)\%$ de confiança, com distribuição t -Student com $n - 2$ graus de liberdade. Um intervalo de confiança para β_0 não é de grande interesse, mas podem ocorrer situações em que seja necessário e para tal e dado por,

$$\left[\hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{Var(\hat{\beta}_0)} \right]$$

Intervalo de Predição para Y_0

Já discutimos anteriormente na secção 3.2.1 como estimamos os valores de Y , agora estamos interessados em encontrar um intervalo de confiança para os valores preditos. De tal forma que

$$\left[\hat{Y}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\sum_{j=1}^n X_j^2 - n\bar{X}^2} \right)} \right]$$

onde \hat{Y}_0 é o valor predito de Y usando o valor particular X ao qual denotamos por X_0 . As inferências podem não ser válidas fora do intervalo de valores de X .

Regressão Linear Múltipla

Intervalo de confiança para β_j

$$\left[E[\hat{\beta}_j] \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 C_{jj}} \right]$$

Intervalo de confiança para $E[\hat{Y}_0]$ e \hat{Y}_0

Intervalos de confiança a $(1 - \alpha)100\%$ para valor esperado no ponto X_0

$$\left[E[\hat{Y}_0] \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 \mathbf{X}_0^T \mathbf{C} \mathbf{X}_0} \right]$$

Intervalos de confiança a $(1 - \alpha)100\%$ para valor ajustado no ponto X_0

$$\left[\hat{Y}_0 \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 [1 + \mathbf{X}_0^T \mathbf{C} \mathbf{X}_0]} \right]$$

MODELOS SÉRIES TEMPORAIS

Neste capítulo serão apresentados os conceitos básicos, métodos e modelos de séries temporais utilizados neste trabalho, seguindo [Bowerman and O'Connell \(1993\)](#), [Pankratz \(1991\)](#), [Hamilton \(1994\)](#) e [Menezes \(2015\)](#).

Segundo [Bowerman and O'Connell \(1993\)](#) uma série temporal (ST) é uma sequência ordenada de observações de uma certa variável.

Por [Hamilton \(1994\)](#), normalmente estas observações tem um início ($t = 1$) e um fim ($t = T$):

$$(Y_1, Y_2, \dots, Y_T)$$

Podemos muitas vezes querer obter observações anteriores ($Y_0, Y_{t-1}, Y_{t-2}, \dots$) ou observações futuras (Y_{T+1}, Y_{T+2}, \dots). A amostra observada (Y_1, Y_2, \dots, Y_T) podia então ser vista como um segmento finito de uma sequência infinita, denotado por $(Y_t)_{t=-\infty}^{\infty}$.

$$(Y_t)_{t=-\infty}^{\infty} = (\dots, Y_{t-1}, \underbrace{Y_0, Y_1, Y_2, \dots, Y_T}_{\text{amostra observada}}, Y_{T+1}, Y_{T+2}, \dots)$$

O estudo da relação entre dados ou observações ordenadas ao longo do tempo é chamado *análise de séries temporais*. ([Pankratz, 1991](#))

No entanto alguns conceitos sobre as componentes de uma Série Temporal são necessários para a compreensão do estudo que se segue. As componentes são:

1. Tendência;
2. Ciclo;
3. Variações Sazonais;
4. Flutuações Irregulares.

Tendência refere-se ao movimento crescente ou decrescente que caracteriza a série temporal num certo período de tempo. Estes movimentos são representativos de vários factores. Por exemplo, a venda de um determinado produto de uma certa empresa num longo período de tempo podem ser influenciados pelas alterações tecnológicas ou mesmo os gostos das pessoas.

Ciclo refere-se a movimentos recorrentes, altos e baixos, à volta de níveis de tendência. Estas flutuações podem ter uma duração entre dois a dez anos, ou ainda maiores quando medidas entre picos.

Variações Sazonais são padrões periodicos numa ST que normalmente se repetem de ano para ano. Estas variações sazonais são usualmente causadas por certos factores como clima e hábitos. Temos, por exemplo, a temperatura média mensal pois temos meses que têm uma certa temperatura média e em outros meses a temperatura média pode ser mais alta ou mais baixa que nesse mês (exemplo: verão e o inverno).

Flutuações Irregulares são movimentos bruscos que ocorrem sem aviso prévio numa ST e não seguem uma padrão que é fácil ser identificado. Esses movimentos representam o que "resta" de uma ST depois da Tendência, Ciclo e Variações Sazonais terem sido considerados. Temos, por exemplo, venda/compra repentina de acções, terremotos, etc.

De salientar que as componentes de uma série temporal nem sempre ocorrem sozinhas, podem ocorrer duas ao mesmo tempo ou até mesmo todas. Por isso não existe um só modelo de previsão, ou seja, nunca iremos obter um modelo que seja "aquele" que melhor representa os nossos dados. (Bowerman and O'Connell, 1993)

4.1 NOTAÇÃO

Nesta secção iremos abordar a notação que irá ser utilizada ao longo deste capítulo. Para tal segue-se o seguinte:

- Y_t - Refere-se à nossa variável resposta, que no caso deste trabalho irá ser referente ao número de admissões/pacientes/utentes diária(o)s no SUHB. Que irá depender de uma ou mais variáveis ou eventos.
- $X_{p,t}$ - Refere-se à variável ou variáveis que podem afetar Y . Temos $X_{p,i}$ variáveis em que $t = 1, \dots, T$ e $T \in \mathbb{N}$. Neste trabalho iremos ver que as variáveis X_i representam respetivamente: uma variável binária que indica se é um dia de feriado ou não, uma

variável com sete níveis que indica o dia da semana (segunda-feira,...) e uma variável com doze níveis que indica o mês do ano (Janeiro,...).

NOTA: apesar de ser usada a notação $X_{p,t}$ para representar as variáveis explicativas, irá haver necessidade de alterar essa notação para $X_{k,t}$ (onde k é o número de variáveis explicativas).

- t - É utilizado como um indicador de tempo: Y_t e X_t refere-se a variáveis cujos valores estão ordenados no tempo. Temos que $t = 1, 2, 3, \dots$, onde t indica o valor da observação correspondente (pode ser valor existente ou valor possível) na sequência dos nossos dados. Ou seja, Y_1 corresponde à primeira observação da variável Y , e podemos continuar assim até ao último valor de t .
- Por vezes é necessário escrevermos o indicador de tempo t como $t - 1, t - 2, t - 3, \dots$ ou como $t + 1, t + 2, t + 3, \dots$. O sinal menos e o sinal mais indicam um tempo anterior ou posterior ao tempo t , respectivamente.

Operadores Lag

O operador lag, ou operador que representa uma variável num intervalo de tempos anteriores, é bastante utilizado em análise de ST isto porque este operador simplifica bastante a notação na escrita de modelos ou processos de uma série temporal. Denotamos este operador por B^i , e quando multiplicamos B^i por uma variável com tempo anterior (ex: Y_{t-1}), t (é indicador de tempo) é movida para trás i vezes. Daí termos,

$$B^i Y_t = Y_{t-i}, \quad B^i \epsilon_t = \epsilon_{t-i}, \quad B^i C = C$$

e uma série que tenha sido diferenciada d vezes é escrita como $(1 - B)^d Y_t$

Função Autocovariância

A função de autocovariância corresponde à intensidade com que pares de valores, com um espaçamento de amplitude k , se acompanham. É dada por:

$$\gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$$

Função Autocorrelação

Em séries temporais estamos interessados no conceito de autocorrelação, isto porque queremos saber o quanto um certo valor da série está relacionado com os seus valores futuros (Y_{t+1}, \dots), ou valores passados (Y_{t-1}, \dots). Autocorrelação mede a direcção e força da

relação entre observações numa série temporal (Y_t) quando as observações são separadas por k períodos, $k = 1, 2, \dots$. Olhamos para Y_t e Y_{t+k} como duas variáveis aleatórias.

O estudo do padrão de autocorrelação em dados temporais levam-nos a identificar o melhor modelo a ajustar para essa série temporal.

A função de autocorrelação, FAC, corresponde à correlação entre pares de valores (Y_t, Y_{t+k}), k é o número de períodos de tempo que separa Y_t de Y_{t+k} . É dada por:

$$\rho_k = \text{Corr}(Y_t, Y_{t+k}) = \frac{\gamma_k}{\gamma_0} = \frac{\text{Cov}(Y_t, Y_{t+k})}{\sigma_Y^2}$$

Cujos coeficientes são estimados a partir da seguinte fórmula:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

Função Autocorrelação Parcial

Outra medida de autocorrelação que nos é muito útil é a autocorrelação parcial. Mede a autocorrelação entre Y_t e Y_{t+k} quando fixamos as variáveis intermédias $Y_{t+1}, Y_{t+2}, \dots, Y_{t+k-1}$.

O conjunto dos coeficientes de autocorrelação parcial num intervalo k é dado por $\{\phi_{kk} : k = 1, 2, \dots\}$, são estimados a partir da seguinte fórmula:

$$\hat{\phi}_{kk} = \begin{cases} \hat{\rho}_1, & \text{se } k = 1. \\ \frac{\hat{\rho}_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} \hat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} \hat{\rho}_j}, & \text{se } k > 1. \end{cases}$$

onde $\hat{\phi}_{kj} = \hat{\rho}_{k-1,j} - \hat{\phi}_{kk} \hat{\rho}_{k-1,k-j}$ $j = 1, 2, \dots, k-1$.

4.2 MODELOS ESTACIONÁRIOS LINEARES

Segundo [Menezes \(2015\)](#): os modelos de séries temporais, também chamados de modelos ARIMA, são modelos que normalmente estão associados a séries temporais e representados por processos estocásticos. Por conseguinte, uma processo estocástico é um conjunto de variáveis aleatórias Y_t ordenadas no tempo.

Dado um processo estocástico Y_t , tal que para todo t temos $E[Y_t^2] < \infty$, então

- $\mu_t = E[Y_t]$, corresponde à função valor médio;
- $\sigma_t^2 = Var(Y_t) = E[(Y_t - \mu_t)^2]$, corresponde à função de variância;
- $\gamma(t_1, t_2) = Cov(Y_{t_1}, Y_{t_2}) = E[(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2})]$, corresponde à função de covariância;
- $\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sigma_{t_1}\sigma_{t_2}} = \frac{Cov(Y_{t_1}, Y_{t_2})}{\sqrt{Var(Y_{t_1})Var(Y_{t_2})}}$, corresponde à função de correlação.

A análise de modelos de séries temporais só é possível se o processo que gera uma série temporal é estacionário. Dizemos então que um processo é estacionário se a média, variância e função de autocorrelação são constantes ao longo do tempo. A nossa noção de estacionaridade é na forma fraca. Também temos a forma forte mas para estar nessa forma a função de distribuição de probabilidade do processo teria de ser todo independente do tempo. De tal forma que, utilizamos a forma fraca. Ou seja,

1. $\mu_t = \mu$;
2. $\sigma_t^2 = \sigma^2$;
3. $Cov(Y_{t_1}, Y_{t_2}) = \gamma(|t_2 - t_1|)$.

Se os dados temporais não forem estacionários, podemos fazer certas modificações para obter uma série estacionária. Uma vez que sabemos quais foram essas modificações na série, podemos mais tarde voltar à série original de forma a comparar os valores que foram previstos.

4.2.1 Modelo White Noise

Ruído Branco ou White Noise, corresponde à sequência $(\epsilon_t)_{t=-\infty}^{\infty}$, cujos elementos tem média zero e variância σ_ϵ^2 ,

1. $E[\epsilon_t] = 0$;
2. $Var(\epsilon_t) = E[\epsilon_t^2] = \sigma_\epsilon^2$;
3. $Cov(\epsilon_{t_1}, \epsilon_{t_2}) = E[\epsilon_{t_1}\epsilon_{t_2}] = 0 \quad t_1 \neq t_2$

ϵ 's não estão correlacionados no tempo.

4.2.2 Processos autoregressivos de ordem 1, $AR(1)$

Um processo autoregressivo de primeira ordem, denotado por $AR(1)$, satisfaz a seguinte equação das diferenças estocásticas:

$$Y_t = c + \phi_1 Y_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0, \sigma^2) \quad (8)$$

onde μ e ϕ_1 são parâmetros a serem estimados, ϵ_t é o termo de erro aleatório e c é uma constante a ser estimada. Este processo (8) diz que qualquer valor de Y_t está relacionado com o valor imediatamente anterior Y_{t-1} através do coeficiente ϕ_1 . ϵ_t é assumido ter média zero e normalmente distribuído (*white noise*). ϵ_t é assumido como independente de Y_{t-1} . O modelo $AR(1)$ pode ser escrito na forma com operador lag

$$(1 - \phi_1 B)Y_t = c + \epsilon_t$$

Existem algumas propriedades deste processo a ter em consideração,

1. Se $|\phi_1| < 1$ temos que Y_t é um processo estacionário.

2. $E[Y_t] = \frac{c}{1-\phi_1}$

3. $\rho_j = \frac{\gamma_j}{\gamma_0} = \phi_1^j$

4. $\gamma_j = \begin{cases} \frac{\sigma^2}{(1-\phi_1^2)}, & \text{se } j = 0. \\ \sigma^2 \frac{\phi_1^j}{(1-\phi_1^2)}, & \text{se } j \geq 1. \end{cases}$

5. $\rho_j = \frac{\gamma_j}{\gamma_0} = \phi_1^j$

4.2.3 Processos autoregressivos de ordem p , $AR(p)$

Um processo autoregressivo de ordem p , denotado por $AR(p)$, é dado por:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t \quad \epsilon_t \sim N(0, \sigma^2) \quad (9)$$

Dado que a média é dada por,

$$\mu = \frac{c}{(1 - \phi_1 - \phi_2 - \dots - \phi_p)} \quad (10)$$

Podemos então escrever a equação (9), usando a equação (10), da seguinte forma:

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \dots + \phi_p (Y_{t-p} - \mu) + \epsilon_t \quad (11)$$

onde $\phi_1, \phi_2, \dots, \phi_p$ são os coeficientes estimados do processo, e ϵ_t ruído branco.

4.2.4 Processos médias móveis de ordem 1, MA(1)

Um processo de médias móveis de ordem 1, denotado por MA(1), é dado por:

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \quad \text{ou} \quad Y_t = \mu + (1 - \theta B)\epsilon_t$$

onde μ e θ são constantes. Seguem as seguintes propriedades

1. $E[Y_t] = \mu$
2. $Var(Y_t) = (1 + \theta^2)\sigma^2$
3. $\gamma_j = \begin{cases} \theta\sigma^2, & \text{se } j = 1. \\ 0, & \text{se } j > 1. \end{cases}$
4. $Corr(Y_t, Y_{t-j}) = \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-j})}} = \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \rho_j, \quad \rho_j = \frac{\gamma_j}{\gamma_0}$
5. $\rho_j = \begin{cases} \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{(1+\theta^2)}, & \text{se } j = 1. \\ 0, & \text{se } j > 1. \end{cases}$

4.2.5 Processos médias móveis de ordem q, MA(q)

Um processo de médias móveis de ordem q, denotado por MA(q), é dado por:

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \dots + \theta_q\epsilon_{t-q}$$

onde ϵ_t é ruído branco e $(\theta_1, \theta_2, \dots, \theta_q)$ são valores quaisquer reais ($\in R$)

1. $E[Y_t] = \mu$
2. $Var(Y_t) = \gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2)\sigma^2$
3. $\gamma_j = \begin{cases} [\theta_j + \theta_{j+1}\theta_1 + \dots + \theta_q\theta_{q-j}]\sigma^2, & \text{para } j = 1, 2, \dots, q. \\ 0, & \text{para } j > q. \end{cases}$
4. A função de autocorrelação é zero para momentos maiores que q.

4.2.6 Processos autoregressivo e médias móveis ARMA(p, q)

Um processo ARMA(p, q) junta termos autoregressivos e de médias móveis num mesmo processo, dado por:

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

onde $\phi_1, \phi_2, \dots, \phi_p$ e $\theta_1, \theta_2, \dots, \theta_q$ são os coeficientes do processo ARMA(p, q).

Significa que um processo ARMA(p, q) tanto está relacionado com valores passados da série como valores passados dos erros.

Comparando a FAC e a FACP conseguimos identificar o processo ARMA a ser utilizado, para isso segue-se a tabela 10.

Tabela 10: Tabela identificação de processo ARMA comparando FAC vs FACP

Processo	FAC	FACP
AR(p)	Decresce para zero exponencialmente ou num padrão de onda ou com ambos	Tem picos até ao tempo p , depois é tudo zero, alguns valores antes do tempo p podem ser zero
MA(q)	Tem picos até ao tempo q , depois é tudo zero, alguns valores antes do tempo q podem ser zero	Decresce para zero exponencialmente ou num padrão de onda ou com ambos
ARMA(p, q)	Decresce para zero exponencialmente ou num padrão de onda ou com ambos	Decresce para zero exponencialmente ou num padrão de onda ou com ambos

4.3 MODELOS NÃO ESTACIONÁRIOS LINEARES

Como foi discutido na secção 4.2 uma série é estacionária quando a média, variância e função de autocorrelação são constantes ao longo do tempo. Caso contrário temos de encontrar métodos para torná-la estacionária, para tal começamos por estabilizar a variância e posteriormente a média se necessário.

4.3.1 Transformação Box-Cox

Métodos para estabilizar a variância devem ser aplicados antes de fazer outra qualquer modificação na ST. Algumas dessas modificações são, por exemplo, pode ser fazer um logaritmo na série ou a raiz quadrada. Existem muitas outras transformações a serem feitas

mas estas duas são as mais utilizadas quando queremos obter uma variância constante, de realçar que os dados devem ser normalmente distribuídos.

A transformação logaritmo e raiz quadrada são ambas membros da família de expoentes por isso são chamadas de transformações de Box-Cox tendo a seguinte equação geral

$$Y'_t = \frac{Y_t^\lambda - 1}{\lambda} \quad (12)$$

onde λ é um número real. Y_t não pode ser negativa. Depois dos dados serem modelados, podemos devolver os valores previstos à escala original.

Reparamos pela equação (12) que se $\lambda = \frac{1}{2}$ temos a transformação raiz quadrada pois $Y_t^{\frac{1}{2}}$ é a raiz de Y_t . Ao subtrairmos 1 e dividindo por λ não altera a estrutura temporal da série, mas sim dá certas propriedades essenciais para que se torne constante.

A escolha mais apropriada para o valor de λ vem da visualização da série, por conseguinte se a variância tende a crescer à medida que o nível da série cresce escolhemos um $\lambda < 1$ caso contrário escolhemos $\lambda > 1$.

4.3.2 Diferenciação

Uma ST que seja estacionária na média é quando esta estabiliza rapidamente numa média constante. Quando isso não ocorre podemos criar uma nova série (W_t) com média constante através de uma diferenciação.

$$W_t = Y_t - Y_{t-1} \quad (13)$$

Caso a série tenha já sofrido uma transformação na variância diferenciamos Y'_t em vez de Y_t . Quando efectuamos as diferenças para todos os valores de t uma vez, chamamos a este processo de primeira diferenciação. Denotamos todas as diferenciações pela letra d , ou seja, se efectuamos uma diferenciação então $d = 1$. Normalmente $d > 2$ nunca chega a ser necessário.

A um processo que tenha sido diferenciado, nós chamamos de processo integrado e daí termos a letra I no acrónimo *ARIMA*. De tal forma que quando fazemos uma diferenciação $d = 1$ temos um processo $ARIMA(p, d, q) = ARIMA(0, 1, 0)$.

Processo *ARIMA*

Resumindo agora o processo $ARIMA(p, d, q)$, este pode ser escrito com recurso ao operador lag da seguinte forma

$\nabla^d = (1 - B)^d$ - corresponde à diferenciação de ordem d

$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)$ - corresponde ao processo AR de ordem p

$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$ - corresponde ao processo MA de ordem q .

Através das formas anteriores podemos definir um processo $ARIMA(p, d, q)$ na sua forma geral:

$$\phi(B)\nabla^d Y_t = c + \theta(B)\epsilon_t$$

4.4 MODELOS DE SÉRIES TEMPORAIS COM COVARIÁVEIS

4.4.1 Modelos de Regressão Dinâmicos

Um modelo de regressão dinâmico, ou modelo DR, mostra como a variável resposta (Y_t) está linearmente relacionado com valores presentes ou passados de uma ou mais variáveis explicativas ($X_{1,t}, X_{2,t}, \dots$).

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \dots + \beta_k X_{k,t} + \epsilon_t \quad n \in N; \quad t = 1, \dots, T \quad T \in R \quad (14)$$

O raciocínio aqui é o mesmo que aquele usado na regressão linear: queremos ajustar um modelo aos dados que melhor representa a relação entre a variável resposta (Y_t) e as variáveis explicativas ($X_{1,t}, X_{2,t}, \dots$) e que estas não são influenciadas pela variável resposta. β_0, β_1, \dots são os coeficientes de regressão. Depois desta relação estar estabelecida podemos proceder a prever valores futuros.

Tendo em conta a equação (14) tem-se um modelo de regressão linear múltipla que também chamamos modelo DR com p variáveis explicativa onde $\beta_0, \beta_1, \dots, \beta_k$ são parâmetros a serem estimados e ϵ_t é o termo de distúrbio/erro estocástico.

Este distúrbio está presente porque não podemos esperar que toda a variação de entrada dada pelas variáveis explicativas consigam explicar toda a variação na variável resposta. Assumimos ter média zero e normalmente distribuído, ao que chamamos *white noise* - variáveis independentes e identicamente distribuídas.

Segundo Pankratz (1991) um possível problema no uso regressão pelo método de mínimos quadrados é que o termo de distúrbio (ϵ_t) do modelo pode estar autocorrelacionado, que significa que ϵ_t pode estar de alguma forma relacionado com os seus valores passados $\epsilon_{t-1}, \epsilon_{t-2}, \dots$, daí termos de efectuar a estimação por máxima verosimilhança. Por exemplo,

pela equação (14) e supondo que ϵ_t está autocorrelacionado e dado que o seu comportamento é descrito por um modelo ARIMA (ex: ARIMA(1,0,2)) temos que:

$$\epsilon_t = \phi_1\epsilon_{t-1} - \theta_1v_{t-1} - \theta_2v_{t-2} + v_t \sim N(0, \sigma_v^2) \quad (15)$$

ou, utilizando os operados lag,

$$(1 - \phi_1B)\epsilon_t = (1 - \theta_1B - \theta_2B^2)v_t \quad (16)$$

Se for feita uma estimação dos coeficientes por mínimos quadrados quando o termo de distúrbio está autocorrelacionado podem ocorrer certos problemas:

1. O estimador de cada parâmetro não é o melhor;
2. A variância dos resíduos subestimar o valor real de σ_ϵ^2 ;
3. o valor de R_a^2 será sobrestimado;
4. os valores previstos pelo modelo vão ser menos precisos uma vez que o padrão da autocorrelação dos resíduos ($\hat{\epsilon}_t$) contém informações importantes que irão estar a ser ignoradas.

De forma a resolver estas situações resolve-se a equação (16) em ordem a ϵ_t ,

$$\epsilon_t = \frac{1 - \theta_1B - \theta_2B^2}{1 - \phi_1B}v_t$$

substitui-se este resultado na equação (14) obtemos a regressão dinâmica

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \frac{1 - \theta_1B - \theta_2B^2}{1 - \phi_1B}v_t$$

ou ainda de forma equivalente podemos utilizar a equação (15) e substituir na equação (14) obtendo-se:

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \phi_1\epsilon_{t-1} - \theta_1v_{t-1} - \theta_2v_{t-2} + v_t \quad (17)$$

onde v_t têm distribuição Normal com média zero e variância constante. Resolvendo a equação (17) em ordem a v_t temos

$$v_t = Y_t - \beta_0 - \beta_1X_{1,t} - \beta_2X_{2,t} - \dots - \beta_kX_{k,t} - \phi_1\epsilon_{t-1} + \theta_1v_{t-1} + \theta_2v_{t-2}$$

onde v_t (\hat{v}_t no caso de uma amostra) é a diferença entre o que observamos (Y_t) e os valores preditos do modelo DR ($\beta_0 - \beta_1X_{1,t} - \beta_2X_{2,t} - \dots - \beta_nX_{n,t} - \phi_1\epsilon_{t-1} + \theta_1v_{t-1} + \theta_2v_{t-2}$). \hat{v}_t representa o conjunto de resíduos se prevermos Y_t usando o modelo DR completo.

Resumindo o que foi discutido nesta secção: se tivermos um modelo DR em que o termo ϵ_t está autocorrelacionado e pode ser descrito por um processo ARIMA(p,d,q). Utilizando a notação descrita no Capítulo 3 obtem-se a seguinte fórmula:

$$\phi(B)\nabla^d\epsilon_t = \theta(B)v_t$$

Resolvendo em ordem a ϵ_t , temos

$$\epsilon_t = \frac{\theta(B)}{\phi(B)\nabla^d}v_t \quad (18)$$

Por último substituindo a equação (18) no modelo representado em (14) obtemos o modelo de regressão em que o termo de distúrbio ϵ_t está autocorrelacionado e representado por um padrão temporal - ARIMA(p,d,q).

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \frac{\theta(B)}{\phi(B)\nabla^d}v_t \quad v_t \sim N(0, \sigma_v^2) \quad (19)$$

No contexto deste trabalho, irá ser aplicada esta metodologia proposta por [Pankratz \(1991\)](#). No entanto, um dos modelos propostos no Capítulo 5 sofre algumas alterações que aquele mencionado em (19).

Essa alteração consiste em utilizar o valor real observado (Y_t) no tempo t subtraíndo-lhe o valor estimado (\hat{Y}_t) pelo modelo nesse mesmo tempo t , em cada um dos termos dados por um processo $AR(p)$. Consequentemente, um modelo DR que sofre tais alterações é dado pela equação

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \dots + \phi_p\epsilon_{t-p} + v_t \quad (20)$$

em que $\epsilon_{t-1} = (Y_{t-1} - \hat{Y}_{t-1})$, $\epsilon_{t-2} = (Y_{t-2} - \hat{Y}_{t-2})$, ..., $\epsilon_{t-p} = (Y_{t-p} - \hat{Y}_{t-p})$, substituindo estes resultados em 20, temos

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \phi_1(Y_{t-1} - \hat{Y}_{t-1}) + \phi_2(Y_{t-2} - \hat{Y}_{t-2}) + \dots + \phi_p(Y_{t-p} - \hat{Y}_{t-p}) + v_t \quad v_t \sim N(0, \sigma_v^2)$$

Utilizando o mesmo exemplo que anteriormente, se estivermos perante um processo ARIMA(p,d,q) (ex: ARIMA(1,0,2)) a equação DR é dada por

$$Y_t = \beta_0 + \beta_1X_{1,t} + \beta_2X_{2,t} + \dots + \beta_kX_{k,t} + \phi_1(Y_{t-1} - \hat{Y}_{t-1}) - \theta_1v_{t-1} - \theta_2v_{t-2} + v_t \quad v_t \sim N(0, \sigma_v^2)$$

onde o termo erro do modelo é dado por

$$\epsilon_t = (1 - \theta_1 B - \theta_2 B^2)v_t + \phi_1(Y_{t-1} - \hat{Y}_{t-1})$$

Analogamente um modelo de regressão em que o termo de erro está autocorrelacionado, pode ser definido por um modelo DR (metodologia de [Pankratz \(1991\)](#)). Outra forma de definirmos estes modelos e estando perante um termo de erro autocorrelacionado (que pode ser descrito por um processo ARIMA(p,d,q)), este pode ser escrito na seguinte forma geral:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \frac{\theta(B)}{\nabla^d} v_t +$$

$$\phi_1(Y_{t-1} - \hat{Y}_{t-1}) + \cdots + \phi_p(Y_{t-p} - \hat{Y}_{t-p}) \quad v_t \sim N(0, \sigma_v^2)$$

MODELAÇÃO DOS DADOS DO SERVIÇO DE URGÊNCIA DO HOSPITAL DE BRAGA

Neste trabalho é estudado os dados do serviço de urgência do Hospital de Braga. Pretendemos prever o número de pacientes que são admitidos diariamente neste serviço.

Tentamos perceber a relação do número de admissões diárias (Y_t) com o tempo (t_t), dia da semana ($Segunda_t, Terça_t, \dots$), mês ($Janeiro_t, Fevereiro_t, \dots$) e os dias de feriado ($Feriado_t, NãoFeriado_t$).

A Figura 13 é o gráfico do número de admissões diárias no SUHB, desde Janeiro 2012 até Dezembro 2016 com um total de 1827 observações. Iremos utilizar estas observações para construir um modelo, e utilizar as últimas 31 observações para verificar a previsão.

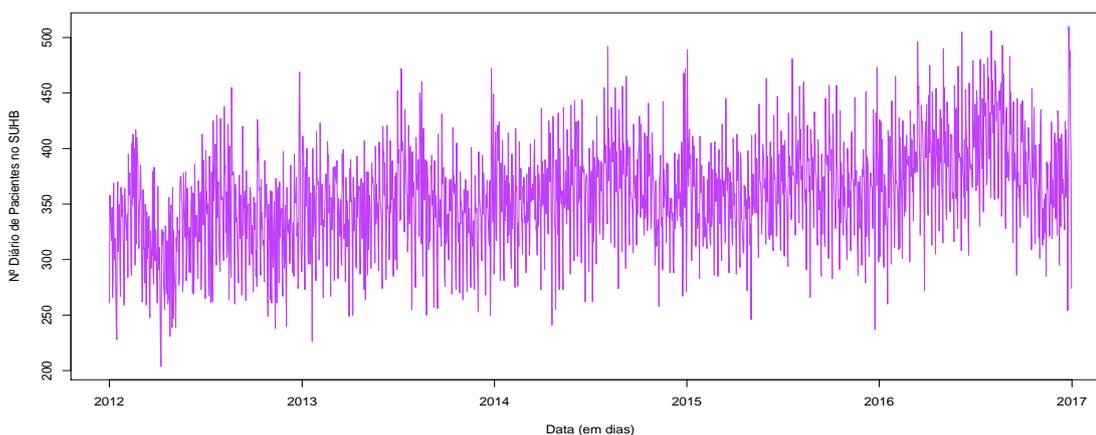


Figura 13: Gráfico do nº admissões diárias no SUHB, Janeiro 2012-Dezembro de 2016

Verificamos que a variância da série é maior em níveis maiores da série, portanto irá ser necessário fazer uma transformação para estabilizar a variância. Então a variável resposta passa a ser ($Y'_t = \frac{(Y_t - 1)^\lambda}{\lambda}$) onde $\lambda = 0.668572$.

5.1 ANÁLISE MODELO DO SUHB

5.1.1 Especificação do Modelo

O modelo do SUHB foi definido pela seguinte equação:

$$\begin{aligned}
 Y'_t = & \beta_0 + \beta_1 t + \beta_2(\text{Feriado}) + \beta_3(\text{Terça}) + \beta_4(\text{Quarta}) + \beta_5(\text{Quinta}) + \beta_6(\text{Sexta}) + \\
 & \beta_7(\text{Sábado}) + \beta_8(\text{Domingo}) + \beta_9(\text{Fevereiro}) + \beta_{10}(\text{Março}) + \beta_{11}(\text{Abril}) + \beta_{12}(\text{Maio}) + \\
 & \beta_{13}(\text{Junho}) + \beta_{14}(\text{Julho}) + \beta_{15}(\text{Agosto}) + \beta_{16}(\text{Setembro}) + \beta_{17}(\text{Outubro}) + \\
 & \beta_{18}(\text{Novembro}) + \beta_{19}(\text{Dezembro}) + \epsilon_t
 \end{aligned}$$

As linhas horizontais, representadas na figura (14), indicam os limites do intervalo de confiança a 95%. É possível observar que a FAC decresce para zero exponencialmente com um padrão de onda enquanto que a FACP tem picos até ao tempo/*lag* sete, depois é tudo zero. O comportamento da FAC e FACP é sugestivo de um processo AR(7), que significa que as observações de cada dia estão correlacionadas com as dos sete dias anteriores. No contexto do nosso problema em estudo significa que o número de pacientes admitidos num dia t está dependente do número de pacientes admitidos nos sete dias anteriores.

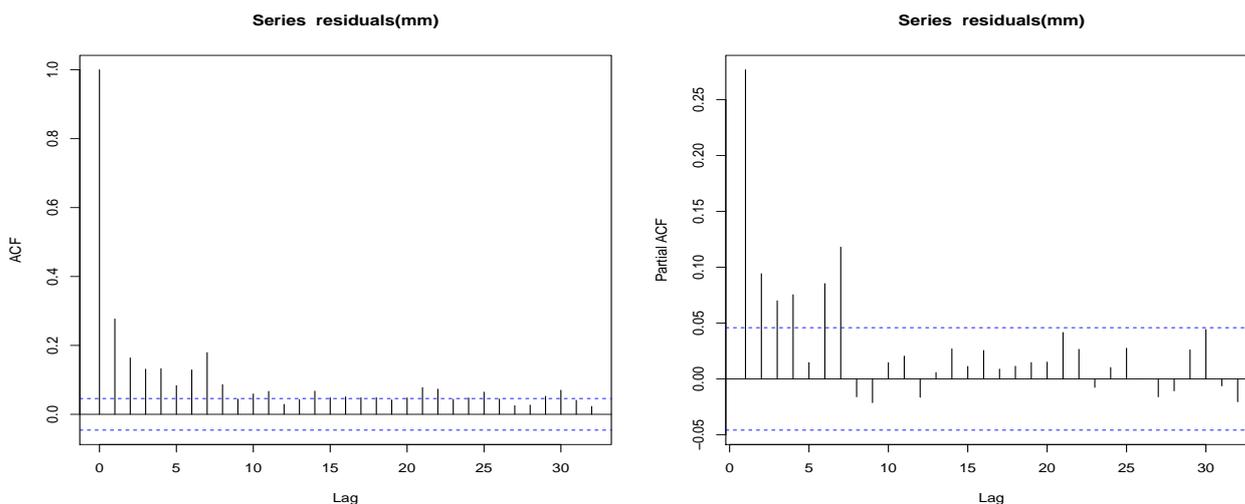


Figura 14: Função de autocorrelação e de autocorrelação parcial dos resíduos do modelo

Outra forma de encontrar um processo ARIMA para explicar o comportamento do distúrbio temporal/termo de erro representado no modelo é recorrendo à função *auto.arima()* do *package* *forecast*. O processo sugerido é ARIMA(2,0,1) ou ARMA(2,1) que por sua vez significa que o número de pacientes admitidos num dia t está dependente do número de pacientes admitidos nos dois dias anteriores, assim como, do erro cometido no tempo t e no dia imediatamente anterior ($t - 1$).

A figura 15 representa a ACF e PACF dos processos AR(7) e ARMA(2,1). Comparando a ACF dos dois processos (figura 15, (a) e (c)) observamos que em (a) todos os valores de autocorrelação encontram-se dentro das linhas do intervalo de confiança de que autocorrelação é zero, enquanto que, em (c) observamos no tempo/lag sete o valor de autocorrelação nesse tempo ultrapassa os limites do intervalo.

Por sua vez, comparando a PACF dos mesmos dois processos (figura 15, (b) e (d)) é possível verificar que em (d) o valor de autocorrelação parcial no tempo/lag 7 ultrapassa os limites do intervalo. Já em (b) no tempo/lag trinta o valor de autocorrelação parcial ultrapassa minimamente o intervalo, no entanto, assumimos como se estivesse dentro.

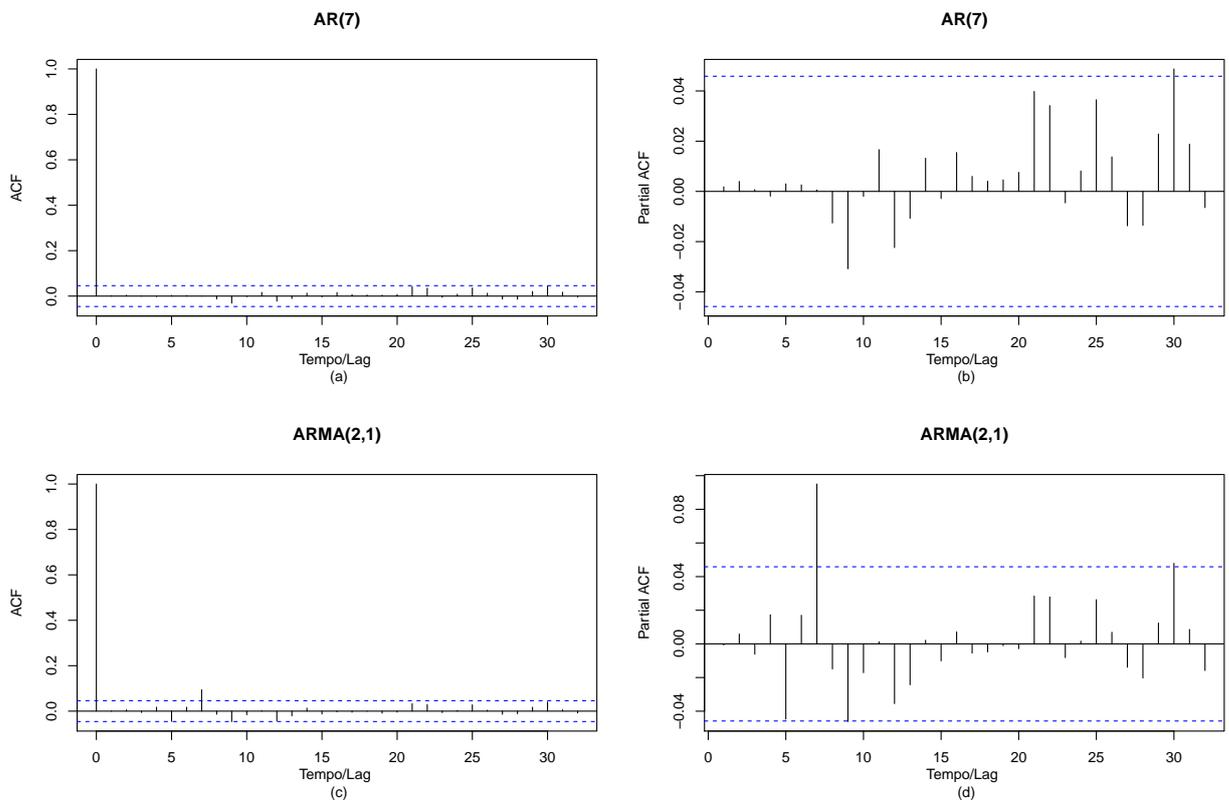


Figura 15: Função de autocorrelação e de autocorrelação parcial dos resíduos dos processo AR(7) e ARMA(2,1)

A tabela 11 identifica os dois modelos ARIMA sugeridos e o seu critério de selecção. Estes critérios confirmam o modelo identificado anteriormente - AR(7) - como o melhor baseado no menor valor de AIC, AICc e BIC.

Concluindo, o modelo autoregressivo de ordem sete é o que melhor representa o termo de erro do modelo.

Tabela 11: Tabela dos diferentes critérios de selecção dos modelos ARIMA sugeridos

Modelos	AIC	AICc	BIC	log likelihood (log verosimilhança)
AR(7)	10005.31	10005.4	10054.9	-4993.65
ARMA(2,1)	10019.57	10019.61	10047.13	-5004.79

O termo de erro do modelo vai ser representado por um processo AR(7), dado pela seguinte equação:

$$\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_3\epsilon_{t-3} + \phi_4\epsilon_{t-4} + \phi_5\epsilon_{t-5} + \phi_6\epsilon_{t-6} + \phi_7\epsilon_{t-7} + v_t \quad v_t \sim N(0, \sigma_v^2)$$

Diagnósticos do Modelo

Nesta secção iremos verificar os pressupostos do modelo AR(7): normalidade, independência, variância constante e média zero.

A figura 16 representa o gráfico da série temporal dos resíduos padronizados do modelo estimado AR(7) referente aos dados do SUHB. O modelo foi ajustado usando método de máxima verosimilhança. Alguns dos resíduos ultrapassam o valor 2. No entanto, os resíduos padronizados parecem estar apresentados de forma aleatória e sem qualquer tipo de padrão aparente à volta da média zero. Através da função de autocorrelação do termo de erro do modelo AR(7), figura 15 (a), podemos afirmar que os erros são independentes.

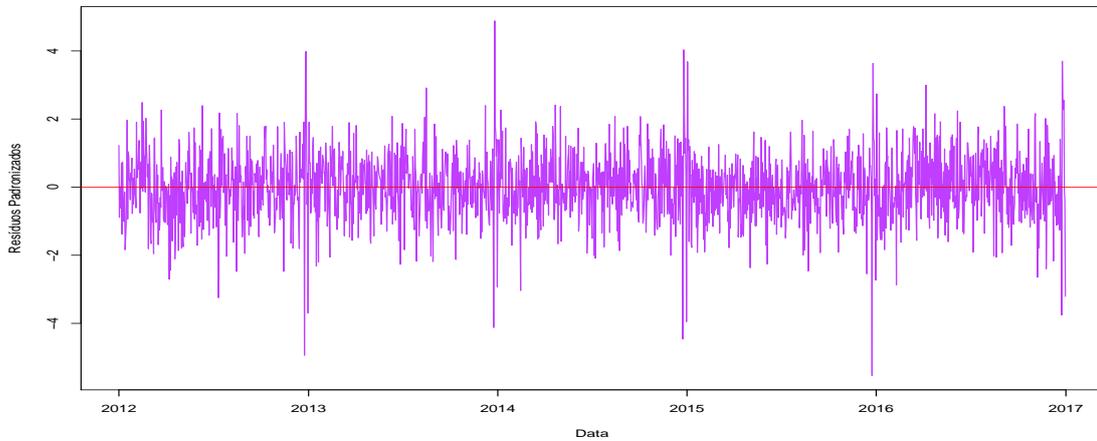


Figura 16: Gráfico dos Resíduos Padronizados ajustados pelo modelo AR(7) aos dados do SUHB

Para verificar a normalidade representamos o gráfico quantil-quantil e o histograma (figura 17) dos resíduos do modelo AR(7) ajustado para os dados do SUHB. No gráfico quantil-quantil observamos que os pontos estão coincidentes com a recta com exceção de alguns pontos nas caudas superior e inferior, o que nos leva a não rejeitar a normalidade do termo de erro do modelo. O Histograma dos resíduos padronizados parecem seguir uma distribuição normal. Outra forma de verificar a normalidade seria realizar um teste de Shapiro-Wilks.

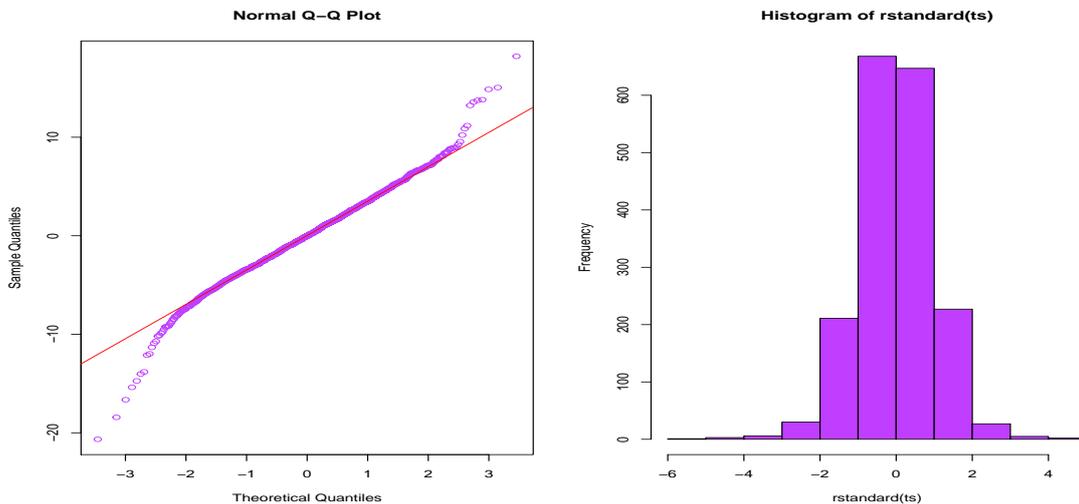


Figura 17: Gráfico Normal Quantil-Quantil e do Histograma dos resíduos obtidos pelo modelo AR(7)

O último pressuposto do modelo é verificar se a variância permanece constante, para tal, representamos o gráfico dos resíduos padronizados vs valores ajustados pelo modelo (\hat{Y}).

Pela Figura 18 observamos que os pontos estão espalhados aleatoriamente e não verificamos um aumento à medida que se aproxima dos valores finais ajustados pelo modelo. Por outras palavras a variância permanece constante.

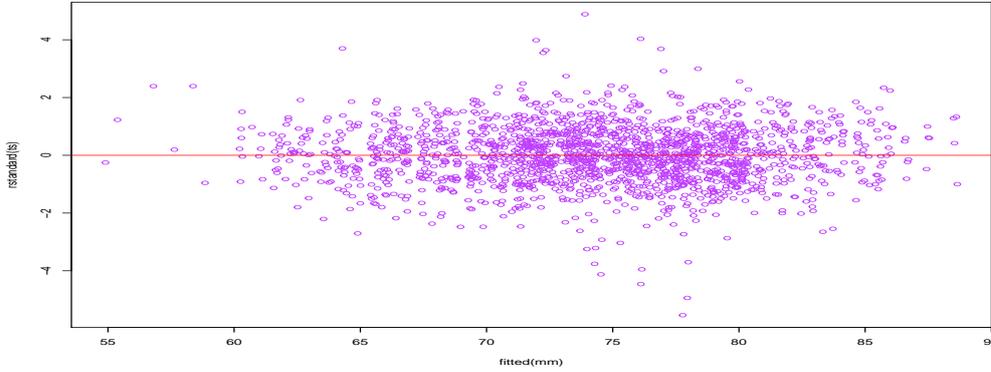


Figura 18: Gráfico resíduos padronizados vs valores ajustados pelo modelo

5.1.2 Modelo Final e Previsão

Uma vez verificado todos os pressupostos do modelo é possível proceder à previsão. No entanto, é necessário definir o modelo final e as estimativas dos parâmetros. O modelo final é dado por

$$\begin{aligned}
 Y'_t = & \beta_0 + \beta_1 t_t + \beta_2(\text{Feriado})_t + \beta_3(\text{Terça})_t + \beta_4(\text{Quarta})_t + \beta_5(\text{Quinta})_t + \beta_6(\text{Sexta})_t + \\
 & \beta_7(\text{Sábado})_t + \beta_8(\text{Domingo})_t + \beta_9(\text{Fevereiro})_t + \beta_{10}(\text{Março})_t + \beta_{11}(\text{Abril})_t + \beta_{12}(\text{Maio})_t + \\
 & \beta_{13}(\text{Junho})_t + \beta_{14}(\text{Julho})_t + \beta_{15}(\text{Agosto})_t + \beta_{16}(\text{Setembro})_t + \beta_{17}(\text{Outubro})_t + \\
 & \beta_{18}(\text{Novembro})_t + \beta_{19}(\text{Dezembro})_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \phi_3 \epsilon_{t-3} + \phi_4 \epsilon_{t-4} + \phi_5 \epsilon_{t-5} + \\
 & \phi_6 \epsilon_{t-6} + \phi_7 \epsilon_{t-7} + v_t \quad v_t \sim N(0, \sigma_v^2)
 \end{aligned} \tag{21}$$

onde v_t é ruído branco, $\hat{\sigma}_v^2 = 13.85$.

As estimativas dos parâmetros do modelo 21 podem ser observadas nas tabelas 12 e 13.

Tabela 12: Tabela resumo do modelo do SUHB

Coefficiente	Estimativa	Erro Padrão	Estatística Z	p-valor
$\hat{\beta}_0$	76.8123	0.4157	184.77	0.0000
$\hat{\beta}_1$	0.0053	0.0002	29.54	0.0000
$\hat{\beta}_8$	-15.3216	0.3483	-43.99	0.0000
$\hat{\beta}_4$	-5.7219	0.3481	-16.44	0.0000
$\hat{\beta}_5$	-6.0047	0.3483	-17.24	0.0000
$\hat{\beta}_7$	-11.4396	0.3481	-32.87	0.0000
$\hat{\beta}_6$	-5.7404	0.3485	-16.47	0.0000
$\hat{\beta}_3$	-5.3629	0.3483	-15.40	0.0000
$\hat{\beta}_{11}$	-0.9987	0.4562	-2.19	0.0287
$\hat{\beta}_{15}$	2.7836	0.4533	6.14	0.0000
$\hat{\beta}_{19}$	-0.7669	0.4563	-1.68	0.0930
$\hat{\beta}_9$	0.1046	0.4619	0.23	0.8209
$\hat{\beta}_{14}$	1.8624	0.4532	4.11	0.0000
$\hat{\beta}_{13}$	0.5669	0.4562	1.24	0.2142
$\hat{\beta}_{12}$	-0.1300	0.4522	-0.29	0.7738
$\hat{\beta}_{10}$	-0.1665	0.4518	-0.37	0.7126
$\hat{\beta}_{18}$	-2.9170	0.4588	-6.36	0.0000
$\hat{\beta}_{17}$	-0.6489	0.4544	-1.43	0.1535
$\hat{\beta}_{16}$	-0.0845	0.4578	-0.18	0.8536
$\hat{\beta}_2$	-6.1149	0.5413	-11.30	0.0000

Interpretação do Modelo:

1. À medida que avançamos no tempo, o número de admissões diárias no SUHB vai aumentar, em média.
2. Quando comparamos os dias de feriado com os dias que não são feriado, o número de admissões nos dias de feriado diminui em relação aos dias em que não é feriado.
3. Quando comparamos os dias da semana (Terça, Quarta, . . . , Domingo) com a Segunda-Feira, o número de admissões nesses dias é sempre inferior, ou seja, a Segunda-Feira é o dia da semana com mais admissões.
4. O número de admissões nos meses de Fevereiro, Junho, Julho e Agosto aumenta em relação ao mês de Janeiro. Enquanto que os meses de Março, Abril, Maio, Setembro, Outubro, Novembro e Dezembro diminuem.

Tabela 13: Tabela das estimativas dos coeficientes do processo AR(7)

Coefficiente	Estimativa	Erro Padrão
$\hat{\mu}$	0.0006	0.1941
$\hat{\phi}_1$	0.2264	0.0233
$\hat{\phi}_2$	0.0642	0.0239
$\hat{\phi}_3$	0.0375	0.0239
$\hat{\phi}_4$	0.0599	0.0239
$\hat{\phi}_5$	-0.0133	0.0240
$\hat{\phi}_6$	0.0584	0.0239
$\hat{\phi}_7$	0.1193	0.0234

Pelo modelo (21) é possível calcular os valores previstos no intervalo de $t = 1$ até $t = 1827$. A Figura 19 ilustra a comparação dos dados originais da série temporal com os dados previstos pelo modelo, foi utilizado o smooth spline de forma a facilitar a sua visualização.

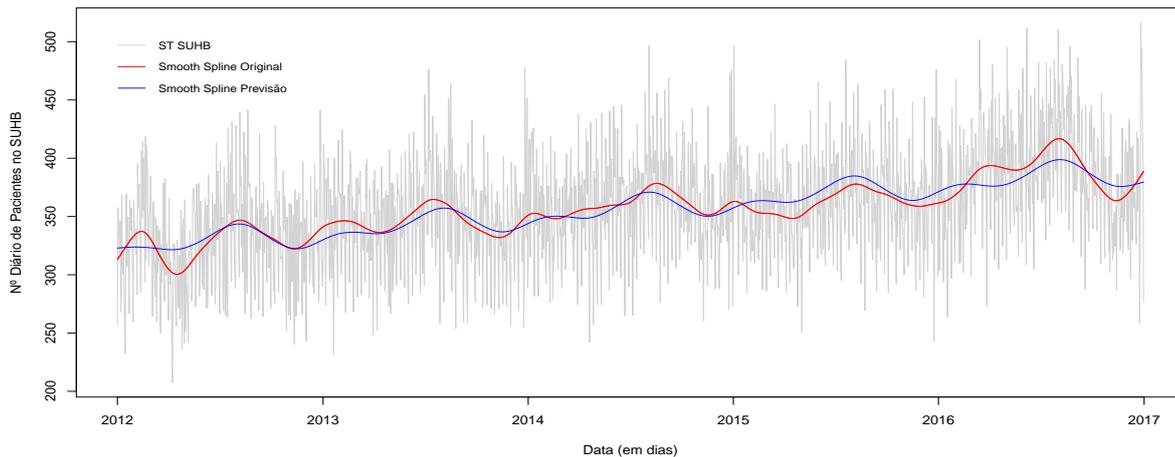


Figura 19: Gráfico de comparação dos dados originais do SUHB vs dados previstos pelo modelo

Iremos proceder à previsão dos valores transformados de Y_t (Y'_t) para o período $t = 1797, \dots, 1827$ (01-12-2016 até 31-12-2016), usando as informações disponíveis até ao tempo $t = 1796$ (30-11-2016). Depois convertemos os valores transformados na unidade de medida original.

A Figura 20 ilustra como a previsão do modelo 21 é realizada para esse período. Os valores de Y_t e \hat{Y}_t encontram-se na métrica original para ser mais fácil a sua compreensão. Ao ser feita a previsão para o tempo $t = 1797$ está-se a somar a parte fixa do modelo (μ_t) ao termo de erro, que está definido por um processo AR(7). Cada termo do processo AR(7) é calculada à custa dos sete termos de erro anteriores, no entanto, quando prevemos para o tempo $t = 1798$ o termo de erro anterior, $\hat{\epsilon}_{1797}$, não foi calculado da mesma forma que os

termos de erro até ao tempo $t = 1796$. Isto acontece porque não conhecemos o valor real de Y_{1797} , daí ser utilizado o valor estimado desse termo de erro (que foi calculado à custa dos 7 termos de erro anteriores). Este processo prosegue desta forma até ao tempo $t = 1827$.

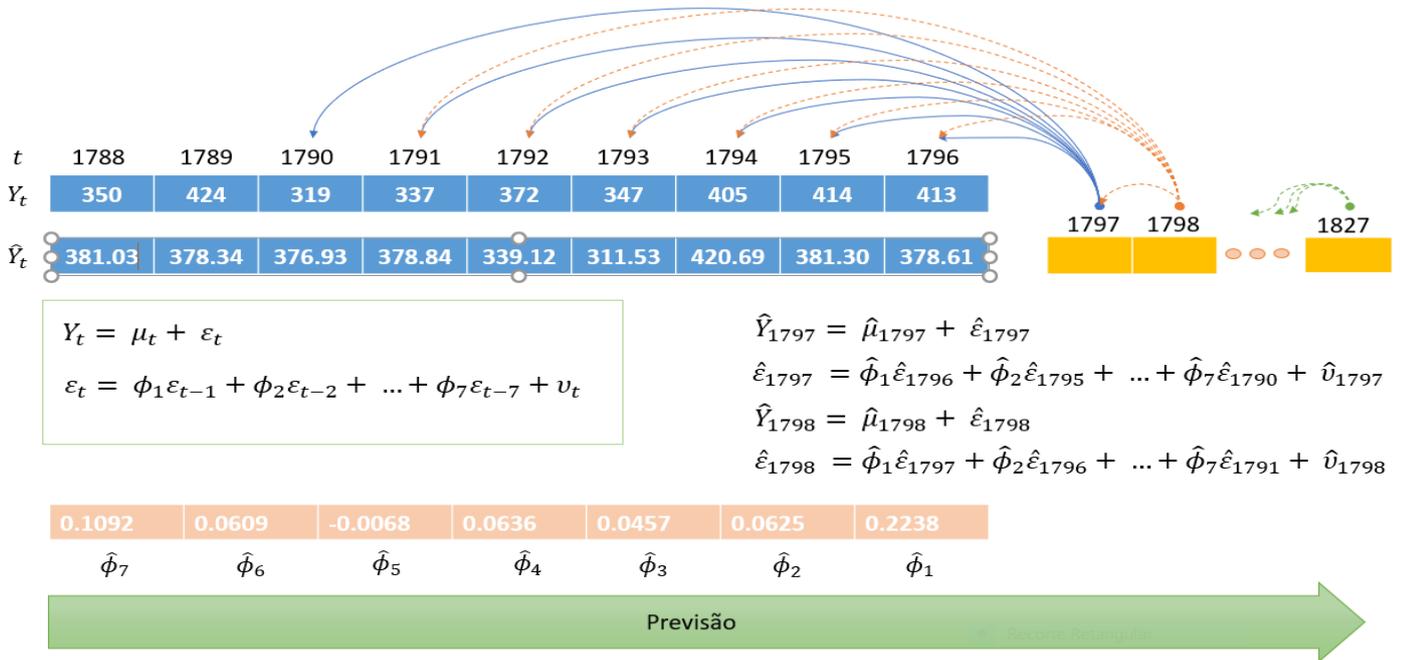


Figura 20: Ilustração de previsão realizada pelo modelo 21

A tabela 14 mostra os valores futuros observados e o erro de previsão (valores observado menos valores previstos). Temos também os limites do intervalo de confinança a 95% à volta do ponto previsto. É possível observar que vários dos pontos observados ($t = 1804, 1820, 1821, 1822, 1823, 1824, 1827$) não se encontram dentro do intervalo de confinça dos correspondentes valores previstos.

Tabela 14: Tabela de Previsão de Y_t obtida do modelo (21) para o mês de Dezembro 2016

Tempo	Data	Limite Inferior	Valor Previsto	Limite Superior	Observado	Erro
1797	2016-12-01	302.30	350.70	402.82	337	-13.70
1798	2016-12-02	342.72	395.57	452.34	405	9.43
1799	2016-12-03	313.54	364.41	419.24	354	-10.41
1800	2016-12-04	284.86	333.53	386.17	322	-11.53
1801	2016-12-05	384.16	440.92	501.74	434	-6.92
1802	2016-12-06	350.90	405.20	463.54	426	20.80
1803	2016-12-07	347.65	401.88	460.18	395	-6.88
1804	2016-12-08	300.74	351.76	406.93	295	-56.76
1805	2016-12-09	343.18	397.83	456.65	409	11.17
1806	2016-12-10	306.20	357.89	413.76	358	0.11

1807	2016-12-11	279.46	328.90	382.51	330	1.10
1808	2016-12-12	382.43	440.27	502.32	413	-27.27
1809	2016-12-13	345.24	400.22	459.41	391	-9.22
1810	2016-12-14	342.44	397.23	456.22	359	-38.23
1811	2016-12-15	340.54	395.21	454.08	372	-23.21
1812	2016-12-16	342.17	396.98	456.00	359	-37.98
1813	2016-12-17	304.53	356.26	412.19	340	-16.26
1814	2016-12-18	278.31	327.78	381.45	327	-0.78
1815	2016-12-19	381.62	439.51	501.64	424	-15.51
1816	2016-12-20	344.17	399.16	458.37	356	-43.16
1817	2016-12-21	341.54	396.32	455.33	417	20.68
1818	2016-12-22	340.10	394.77	453.66	394	-0.77
1819	2016-12-23	341.85	396.66	455.69	374	-22.66
1820	2016-12-24	304.21	355.93	411.86	254	-101.93
1821	2016-12-25	241.11	287.19	337.46	357	69.81
1822	2016-12-26	381.48	439.38	501.51	510	70.62
1823	2016-12-27	344.05	399.04	458.25	473	73.96
1824	2016-12-28	341.47	396.26	455.26	488	91.74
1825	2016-12-29	340.11	394.79	453.68	423	28.21
1826	2016-12-30	341.89	396.72	455.76	394	-2.72
1827	2016-12-31	304.27	356.00	411.94	274	-82.00
Tempo	Data	Limite Inferior	Valor Previsto	Limite Superior	Observado	Erro
MAPE	=0.08295739	= 8.3%				

De forma a visualizar os dados apresentados na tabela 14, construiu-se o gráfico dado pela figura 21 onde é possível visualizar os valores previstos pelo modelo (21) para o mês de Dezembro 2016 e o respectivo intervalo de confiança a 95%. Posteriormente faz-se representação gráfica da comparação dos dados do SUHB com a previsão realizada.

Pela tabela obtemos ainda o erro percentual absoluto médio (MAPE) é de 8.3% < 10%, o que significa que a previsão é bastante precisa quando comparada com os valores reais observados nesse período.

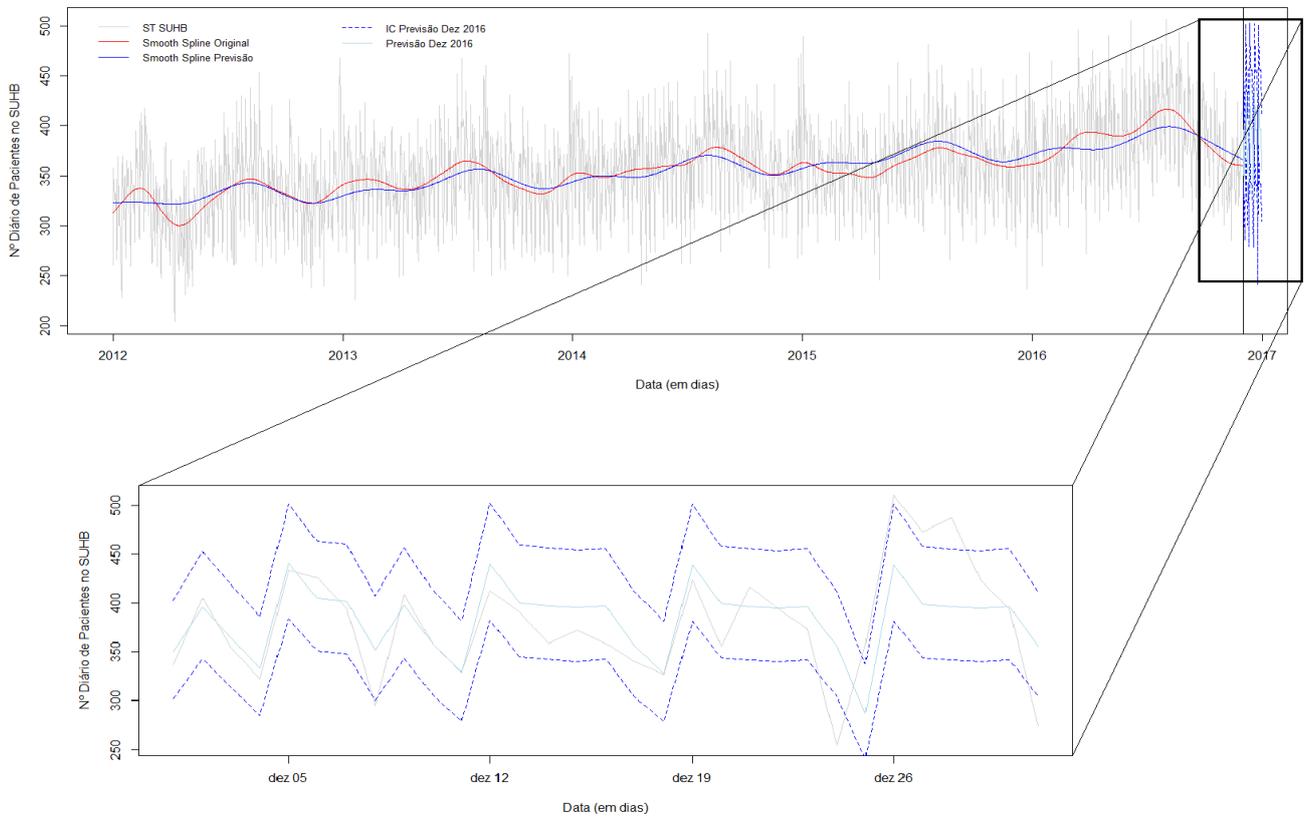


Figura 21: Gráfico de comparação dos dados originais do SUHB vs valores futuros para o mês de Dezembro 2016

Uma possibilidade encontrada para melhorar a previsão é: uma vez que o termo de erro segue um processo $AR(7)$, concluímos que a previsão de um valor futuro (Y_{t+1}) tem de ter em consideração os sete dias anteriores ao dia de previsão. Consequentemente cada termo do processo $AR(7)$ irá ter a seguinte forma $\left[(Y_{(t+1)-1} - \hat{Y}_{(t+1)-1}), \dots, (Y_{(t+1)-p} - \hat{Y}_{(t+1)-p}) \right]$. A equação de previsão dada pelo modelo 21 irá ser dada por,

$$\begin{aligned}
 Y'_{t+1} = & \beta_0 + \beta_1 t_{t+1} + \beta_2 (Feriado)_{t+1} + \beta_3 (Terça)_{t+1} + \beta_4 (Quarta)_{t+1} + \beta_5 (Quinta)_{t+1} + \\
 & \beta_6 (Sexta)_{t+1} + \beta_7 (Sábado)_{t+1} + \beta_8 (Domingo)_{t+1} + \beta_9 (Fevereiro)_{t+1} + \beta_{10} (Março)_{t+1} + \\
 & \beta_{11} (Abril)_{t+1} + \beta_{12} (Maio)_{t+1} + \beta_{13} (Junho)_{t+1} + \beta_{14} (Julho)_{t+1} + \beta_{15} (Agosto)_{t+1} + \\
 & \beta_{16} (Setembro)_{t+1} + \beta_{17} (Outubro)_{t+1} + \beta_{18} (Novembro)_{t+1} + \beta_{19} (Dezembro)_{t+1} +
 \end{aligned}$$

$$\begin{aligned} &\phi_1(Y_{(t+1)-1} - \hat{Y}_{(t+1)-1}) + \phi_2(Y_{(t+1)-2} - \hat{Y}_{(t+1)-2}) + \phi_3(Y_{(t+1)-3} - \hat{Y}_{(t+1)-3}) + \\ &\phi_4(Y_{(t+1)-4} - \hat{Y}_{(t+1)-4}) + \phi_5(Y_{(t+1)-5} - \hat{Y}_{(t+1)-5}) + \phi_6(Y_{(t+1)-6} - \hat{Y}_{(t+1)-6}) + \\ &\phi_7(Y_{(t+1)-7} - \hat{Y}_{(t+1)-7}) + v_{t+1} \quad v_{t+1} \sim N(0, \sigma_v^2) \end{aligned} \quad (22)$$

A Figura 22 ilustra como a previsão é realizada segundo a equação de previsão dada por 22. Os valores Y_t e \hat{Y}_t encontram-se na métrica original para ser mais fácil a sua compreensão. Pretendemos fazer a previsão para o período $t = 1797$ até $t = 1827$ (Dezembro 2016), quando fazemos a previsão para o tempo $t = 1797$ (1 de Dezembro 2016) está-se a somar a parte fixa do modelo (μ_t) ao termo de erro que está definido por um processo AR(7). Cada termo do processo AR(7) é calculado à custa dos 7 termos de erro anteriores, no entanto, quando pretendemos prever para o tempo $t = 1798$ (2 de Dezembro 2016), e segundo a equação de previsão 14, como não é conhecido o valor real de Y_{1797} o que irá ser feito é: o utilizador irá inserir esse valor pondendo-se assim calcular o valor do termo de erro $\hat{\varepsilon}_{1797}$ em vez de fazer a estimação desse termo. A previsão será realizada desta forma até ao último valor $t = 1827$ (31 de Dezembro de 2016).

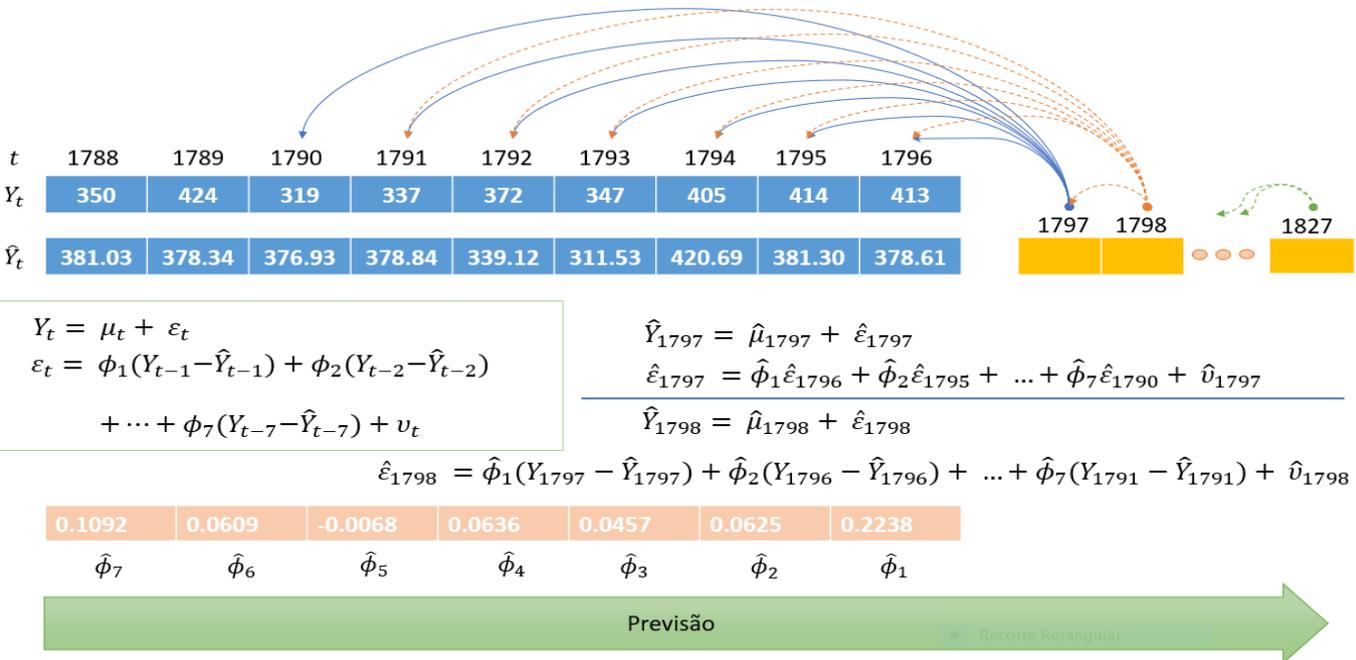


Figura 22: Ilustração de previsão realizada pelo modelo 21

Utilizando a equação de previsão (22) e fazendo a previsão para o mês de Dezembro de 2016 obtemos os dados apresentados na tabela 15.

Tabela 15: Tabela de Previsão de Y_t obtida através da equação (22), para o mês de Dezembro de 2016

Tempo	Data	Valor Previsto	Observado	Erro
1797	2016-12-01	350.67	337	-13.67
1798	2016-12-02	392.25	405	12.75
1799	2016-12-03	365.51	354	-11.51
1800	2016-12-04	331.18	322	-9.18
1801	2016-12-05	436.64	434	-2.64
1802	2016-12-06	403.08	426	22.92
1803	2016-12-07	403.76	395	-8.76
1804	2016-12-08	349.54	295	-54.54
1805	2016-12-09	384.29	409	24.71
1806	2016-12-10	355.58	358	2.42
1807	2016-12-11	324.78	330	5.22
1808	2016-12-12	437.43	413	-24.43
1809	2016-12-13	397.32	391	-6.32
1810	2016-12-14	388.93	359	-29.93
1811	2016-12-15	378.70	372	-6.70
1812	2016-12-16	388.38	359	-29.38
1813	2016-12-17	344.69	340	-4.69
1814	2016-12-18	317.60	327	9.40
1815	2016-12-19	431.38	424	-7.38
1816	2016-12-20	389.28	356	-33.28
1817	2016-12-21	379.06	417	37.94
1818	2016-12-22	391.00	394	3.00
1819	2016-12-23	389.50	374	-15.50
1820	2016-12-24	347.51	254	-93.51
1821	2016-12-25	264.71	357	92.29
1822	2016-12-26	444.16	510	65.84
1823	2016-12-27	407.96	473	65.04
1824	2016-12-28	414.68	488	73.32
1825	2016-12-29	426.36	423	-3.36
1826	2016-12-30	405.53	394	-11.53
1827	2016-12-31	357.01	274	-83.01

$$\text{MAPE} = 0.07802228 = 7.8\%$$

Pela tabela 15, podemos contruir um gráfico (Figura 23) onde representamos o número de admissões diárias previstos diariamente para o mês de Dezembro de 2016 contra o número de admissões diárias que foram realmente observadas nesse período de previsão. Inclui-se no gráfico também a ST das admissões diárias do SUHB até ao final do mês de Novembro 2016, juntamente, encontra-se a suavização dos dados da ST do SUHB e dos valores previsto pelo modelo até essa mesma data.

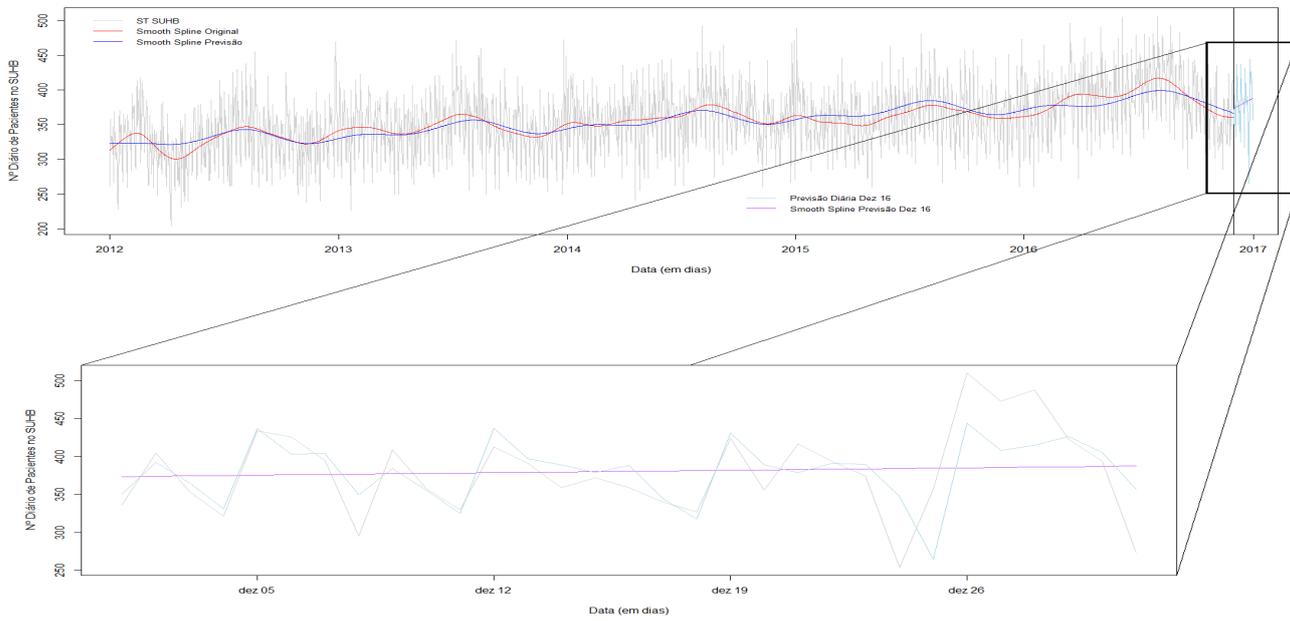


Figura 23: Gráfico de comparação dos dados originais do SUHB vs valores futuros (previstos diariamente) para o mês de Dezembro 2016

APLICAÇÃO SHINY PARA URGÊNCIA DO HOSPITAL DE BRAGA

Shiny é um *package* da biblioteca do software R que torna mais fácil a construção de Web Apps através do R. Pode-se construir apps para serem utilizadas numa página web ou através de um script no R ou ainda a construção de dashboards. Pode-se também melhorar a aplicação em Shiny recorrendo a temas de CSS, Htmlwidgets ou JavaScripts. [RStudio \(2017\)](#)

Porquê criar uma aplicação em Shiny?

Para quem conhece o software R sabe que quando criamos um script e o queremos executar, este processo não é tão fácil para quem nunca utilizou este tipo de linguagem, ou seja, não é user friendly.

Para tal criamos então uma aplicação em Shiny de forma a facilitar a compilação de um script em R, tornando assim possível que pessoas que não percebiam da linguagem em R possam utilizar funções desse script.

Uma das funcionalidades desta aplicação é que pode ser publicada online de forma a estar disponível a todos, fazendo o que o nosso script em R faz mas sem precisarmos de escrever nada, a não ser "clique" num Website.

Uma das muitas vantagens de utilizar Shiny é: juntos R e Shiny têm uma grande característica "reacção", em inglês chamamos *reactive* (para se mais fácil a compreensão iremos utilizar a expressão em inglês no precurso deste capítulo) significa que é possível fazer um update à variável de saída (*output*) à medida que se altera as variáveis de entrada (*input*) através de widgets, ver figura 24. Alguns desses widgets inclui **Date Input** que permite ao utilizador introduzir uma data, **File Input** que permite fazer o upload de base de dados (ficheiros excel), entre muitos outros.

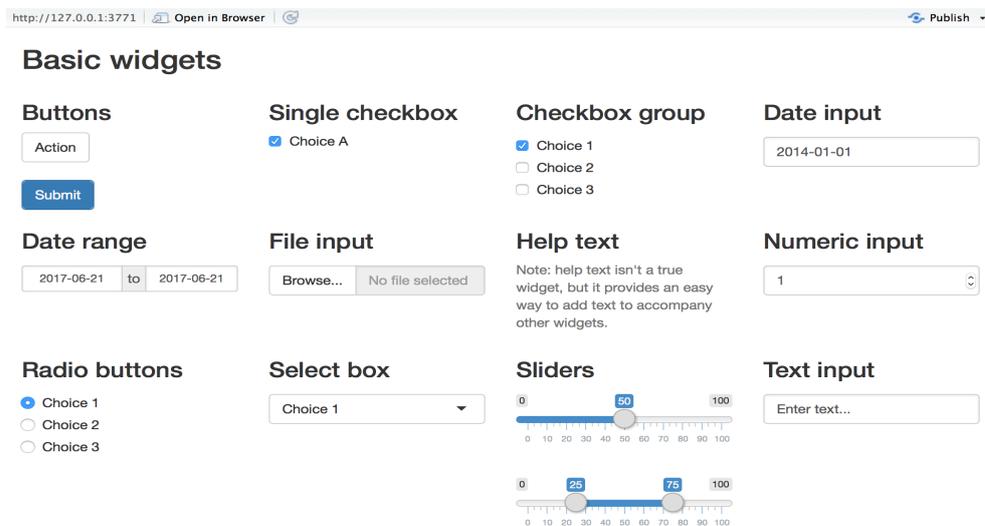


Figura 24: Widgets no package Shiny (<https://shiny.rstudio.com/>)

6.1 ESTRUTURA DE UMA APLICAÇÃO SHINY

Uma aplicação em Shiny pode ter como servidor o nosso próprio computador ou pode estar armazenada numa "cloud" online de forma que todos os utilizadores tenham acesso.

Para construirmos uma aplicação em Shiny temos de ter em conta 2 componentes,

1. Temos um documento Web (HTML) que será o documento que contará com as informações que os utilizadores terão acesso, chamamos a este documento *user interface* (UI). No entanto não precisamos de saber nada sobre HTML uma vez que o R gera esse documento automaticamente.
2. A segunda componente da nossa aplicação é um novo documento que conterà as instruções em R para que o nosso servidor possa seguir. Estas instruções serão necessárias para o servidor saber o que fazer quando o utilizador muda as opções na App.

Todo o código gerado para fazer o display da UI (ver código 6.1) precisa de estar contido entre parênteses de uma função do layout. As funções de layout como `fluidPage` (existem muitas mais disponíveis) criam uma tela branca para a interface e usa a função `fluidRow` para posicionar os Widgets que o utilizador usa (`sliderInput`, `dateInput`,...) e output (`plotOutput`, `textOutput`,...). Cada função de layout têm a sua forma de ser colocada no documento, mas outras como `fixedPage` e `navbarPage` criam diferentes estilos de páginas. No entanto, todas seguem a mesma estrutura hierárquica mas funções elementares como Widgets têm de ser colocados dentro de uma função de layout (`fluidPage`). Funções do

mesmo nível estão separadas por “,”.

Código 6.1: Exemplo de documento UI Layout

```
ui <- fluidPage(
  fileInput(inputId = "dados1", "Insira 1 Base de Dados"),
  helpText("Tamanho máximo 9MB"),
  hr(),
  radioButtons("remove", "Remover 2 Base de Dados:",
    choiceNames = list(
      "Nao",
      "Sim"),
    choiceValues = list(
      "Nao", "Sim"),
    inline = T)
  verbatimTextOutput("oid1")
)
```

O código apresentado em 6.1 mostra um objecto *output* definido por *verbatimTextOutput(inputId = "oid1")* que indica qual será o local do layout onde irá ser apresentado, depois transmite na função *server* que o conteúdo contido nesse objecto de reacção será do tipo texto.

A estrutura do documento *server* (Código 6.2) tem um papel importante pois define as instruções para a aplicação, também reduz significamente o tempo de computação e aumenta a velocidade de reacção da aplicação.

A função, *shinyApp*, necessita de um objecto de entrada (valores da UI) e um objecto de saída (objectos a serem chamados pela função *server*). Os objectos que se alteram dependendo do que foi colocado nos Widgets da UI, como *input\$dados1* no exemplo dado pelo código 6.1, estes termos são de “reacção” ou *reactive*. Sempre que alteramos o input dado pelos widgets, o valor do objecto *reactive* sofre um update para ser observada essa alteração.

As expressões *reactive* usam expressões e funções escritas com código R, que por sua vez usam os widgets para devolverem um objecto, no exemplo apresentado em 6.2, o objecto **data** recebe um *input* da UI que posteriormente utiliza esse *input* para criar um *data frame* com o nome *urg.geral.1* e por fim devolve esse mesmo objecto. E como discutimos anteriormente este objecto **data** pode ser usado dentro de funções *render** ou expressões *reactive*.

Mas, expressões *reactive* não podem enviar os seus objectos para a UI, por isso, para o fazer é necessário a utilização de funções *render** como *renderPrint*.

Código 6.2: Exemplo de documento Server

```
server <- function(input, output, session){
```

```

output$oid1 <- renderPrint({
  tabela=as.data.frame(noquote(strsplit(input$vec1, ",")[[1]]))
  colnames(tabela)<-"Datas_␣Introduzidas:"
})

data <- reactive({
  file1<-input$dados1
  urg.geral.1 <- read.csv2(file=file1$datapath,header=T,
  col.names=c("Data","n.admiss","t","dia.semana","mes","feriados"),
  fileEncoding = "latin1")

  return(urg.geral.1)
})
}

shinyApp(ui = ui, server = server)

```

Expressões que contêm objectos de reacção necessitam estar incluídos numa função *render** (para devolver objectos de reacção à UI, onde * define a descrição desse objecto como plot, table, ou text) ou uma expressão de reacção (que normalmente, são uma série de base de dados que podem ser enviadas para a UI através de funções *render*). Alguns exemplos dos vários tipos de funções *render* que foram utilizadas na construção da app apresentada na secção 6.3 pode ser visto na Figura 25

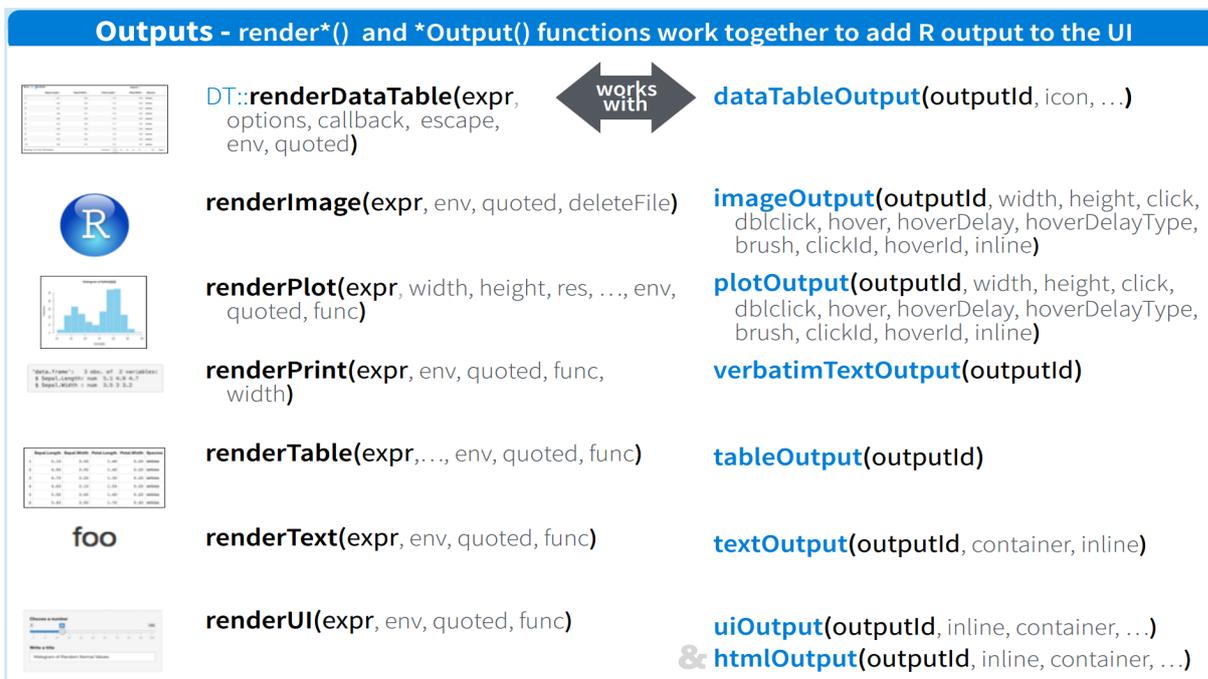


Figura 25: Funções *render* no package Shiny

O código no início do script (documento R) só é compilado uma vez aquando a inicialização da aplicação. Esta área é a mais apropriada para a colocação das livrarias ou *library's* a serem carregadas pelo R, base de dados, objectos com valores fixos que não precisam de serem compilados todas as vezes que o utilizador alterar algum widget, desta forma reduzindo o tempo de computação. (J Wojciechowski and Upton, 2015)

Código 6.3: Exemplo de código a ser colocado no início do script R

```
library(shiny)
library(data.table)
library(forecast)
library(DT)
library(ggplot2)

urg.16.update <- read.csv2("urg.16_update.csv",header = T,
                           fileEncoding = "latin1")
urg.16.update$Data=as.Date(urg.16.update$Data)
```

Muito mais podia ser tratado sobre a construção de uma aplicação através do R e Shiny, no entanto este não é o foco deste trabalho. Para mais informações sobre a construção de uma aplicação em Shiny podem ler alguns dos artigos que foram essenciais para a construção da aplicação que é discutida na Secção 6.3: (Allaire, 2014), (Grolemund, 2017), (Oleś, 2015) e (Beeley, 2013).

6.2 PARTILHAR APLICAÇÃO SHINY

Uma vez criada a nossa aplicação em Shiny, é normal que queiramos partilhá-la, com colegas ou mesmo com outros utilizadores de R. Podemos querer partilhar com pessoas que mesmo não estando por dentro do que é o R possam utilizar a aplicação que criámos. Ou, podemos partilhar com todo o Mundo através de servidores.

Existem várias formas de o fazer:

- Partilhar com utilizadores R, através de GitHub/Gist
- Comprimindo os ficheiros da aplicação num ficheiro .rar ou .zip, e partilhando num serviço FTP como DropBox.
- Partilhar através da Internet usando serviços pagos ou gratuitos como ShinyApps.io
- Ou usando localmente o Web-Browser

6.3 HOSPITAL DE BRAGA: WEB-APP DE PREVISÃO

Nesta secção iremos discutir a web-app que foi construída no desenvolvimento deste trabalho para o Hospital de Braga.

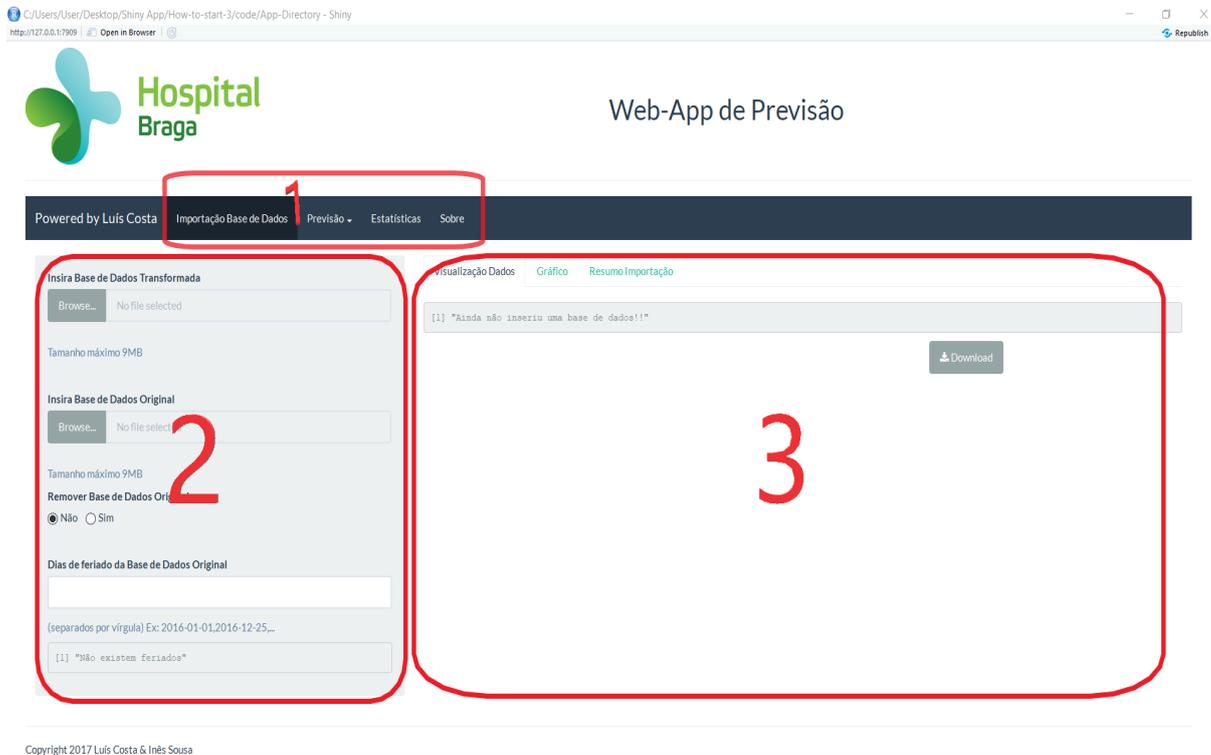


Figura 26: Layout da Web-App de Previsão do SUHB

Na Figura 26 temos o layout (UI) da aplicação que o utilizador tem acesso. Toda a aplicação terá o layout dado pela Figura 26 e os números 1, 2 e 3 representam o seguinte:

1. Representa os separadores que a aplicação contém, ou seja:
 - **Importação Base de Dados.**
 - **Previsão**, que contém ainda mais dois separadores: **Previsão Mensal** e **Previsão Diária.**
 - **Estatísticas.**
 - **Sobre**
2. Representa o *SidebarPanel* que contém todo o tipo de interface que o utilizador terá para interagir com o *output*.
3. Por último, temos o *MainPanel* que será onde todo o *output* da aplicação será visualizado.

6.3.1 Importação Base de Dados

O separador da importação das base de dados é dos mais importantes, porque se não for inserido nenhum ficheiro excell que contém a base de dados a analisar não será possível fazer previsões. A Figura 27 representa a interface que o utilizador tem acesso.

Visualização Dados Gráfico Resumo Importação

Mostrar 10 registos Procurar:

Data	n.admiss	t	dia.semana	mes	feriados
2012-01-01	261	1	domingo	janeiro	feriado
2012-01-02	358	2	segunda-feira	janeiro	nao feriado
2012-01-03	335	3	terça-feira	janeiro	nao feriado
2012-01-04	334	4	quarta-feira	janeiro	nao feriado
2012-01-05	317	5	quinta-feira	janeiro	nao feriado
2012-01-06	347	6	sexta-feira	janeiro	nao feriado
2012-01-07	266	7	sábado	janeiro	nao feriado
2012-01-08	284	8	domingo	janeiro	nao feriado
2012-01-09	369	9	segunda-feira	janeiro	nao feriado
2012-01-10	312	10	terça-feira	janeiro	nao feriado

Mostrando de 1 até 10 de 1,492 registos

Anterior 2 3 4 5 ... 150 Seguinte

Figura 27: Separador "Importação Base de Dados" da Web-App de Previsão do SUHB

Nesta secção irei fazer uma pequena descrição da utilização deste separador. No *Side-barPanel* temos dois *fileinput* onde são inseridos os dois ficheiros excell com os dados, no primeiro inserimos a base de dados transformada (contém as variáveis: Data, n.admiss, dia.semana, mes e feriados). No segundo inserimos um ficheiro excell com a base de dados original (uma linha por paciente).

Caso não queiramos utilizar a base de dados original basta selecionar a opção "Sim", caso contrário selecionar opção "Não". Para inserirmos os dias de feriados contidos na base de dados original basta colocar a data, no formato ano-mes-dia, referente aos dias feriados separadas por ",".

Uma vez preenchido o *sidebarPanel* iremos visualizar no *mainPanel* três separadores:

1. Visualização de Dados;
2. Gráfico;
3. Resumo de Importação.

Em **Visualização de Dados** iremos ter uma tabela referente à base de dados transformada e caso tenhamos inserido uma base de dados original, esta irá ser transformada e unida à base de dados transformada. A tabela tem seis colunas, referente às variáveis: Data, n.admiss, t, dia.semana, mes e feriados. Nesta tabela podemos fazer pesquisas de forma a filtrar o seu conteúdo, como por exemplo, se quisermos só observar as datas de um ano em específico, ou se quisermos visualizar os linhas cujo dia da semana é o domingo, entre outros. Por fim podemos fazer o download deste ficheiro, para isso basta clicar no botão "download" localizado à direita da tabela.

No separador **Gráfico**, como o próprio nome indica podemos visualizar o gráfico referente ao número de admissões diárias no SUHB no período das base de dados inseridas anteriormente. Podemos também fazer o download desse gráfico no formato *.pdf*.

No separador **Resumo de Importação** temos as estatísticas referentes a todas as variáveis observadas na tabela discutida anteriormente. Algumas dessas estatísticas são a média, mediana, 1^o, 3^o e 4^o quantil, valor mínimo e máximo observado.

6.3.2 Previsão

O separador Previsão está dividido em outros dois sub-separadaores:

1. Previsão Mensal;
2. Previsão Diária.

O separador referente à Previsão Mensal (ver Figura 28) é onde podemos fazer a previsão mensal.

No *sidebarPanel* é onde será introduzido o período de previsão, para isso temos dois *input's* de datas um referente à data de início (ex: 2016-01-01) e outro à data final de previsão (ex: 2016-01-31). Depois teremos de adicionar os dias de feriado contidos no período de previsão separados por ",", tal como foi tratado na Secção 6.3.1.

No *mainPanel* iremos ter dois separadores:

1. Tabela Previsão;
2. Gráfico Previsão.

No primeiro separador temos a tabela onde são apresentados todos os valores previstos do número diário de admissões no SUHB. A tabela contém 4 colunas: Data de previsão, Limite Inferior, Valor Previsto e Limite Superior. Também podemos usar a pesquisa para procurar, por exemplo, o valor previsto numa determinada data.

No segundo separador é apresentado o gráfico de previsão, onde observamos uma linha referente aos valores previstos, e outras duas linhas referentes ao limite inferior e limite superior.

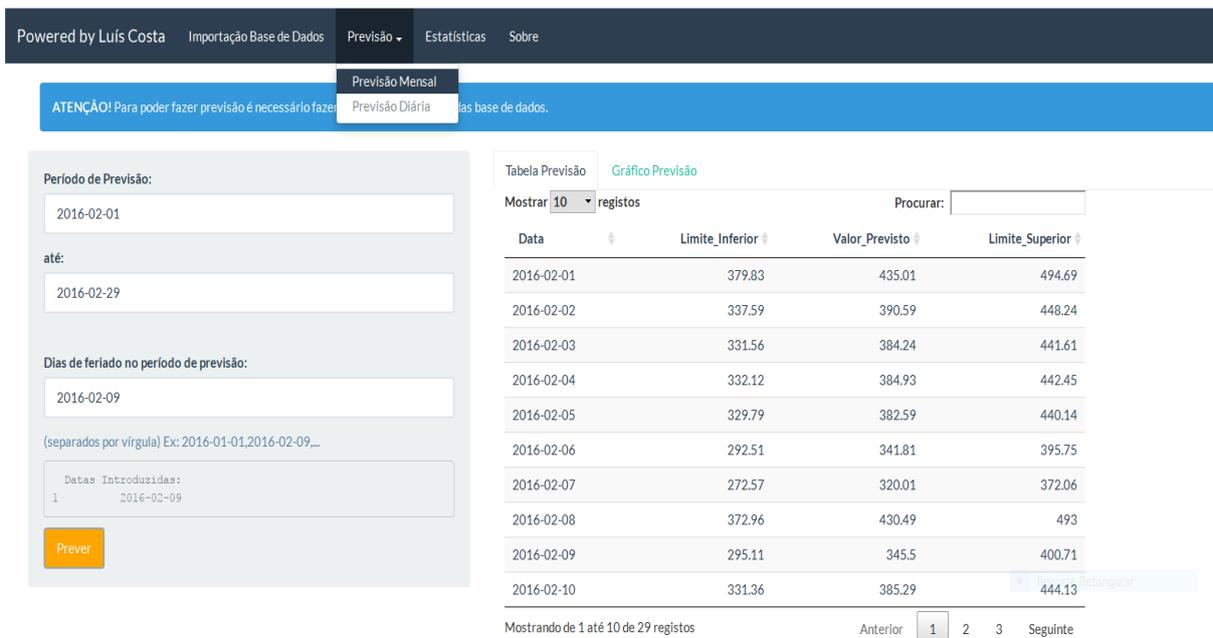


Figura 28: Separador "Previsão Mensal" da Web-App de Previsão do SUHB

Passando ao separador da Previsão Diária (Figura 29), cujo desenvolvimento foi o foco da realização deste trabalho, é possível observar que o *sidebarPanel* tem um layout diferente. Esta diferença deve-se ao facto de a previsão ser feita dia a dia e com base no número de admissões nos 7 dias imediatamente anteriores ao dia de previsão.

No *sidebarPanel* temos apenas um *input* para o dia de previsão (ex: 2016-02-01) de seguida é perguntado se esse dia de previsão é feriado, em caso afirmativo seleccionamos a opção "Sim" e "Não" caso contrário.

De forma a introduzirmos o número de admissões dos 7 dias anteriores ao dia de previsão foi colocado 7 *numericInput* com o dia corresponde de cada *input*. Na Figura 29 temos 7 "caixas" com os valores: 443, 399,385,373,356,311,337 que representam o número diário de admissões no SUHB no período 2016-01-25 até 2016-01-31 (7 dias anteriores ao dia de previsão 2016-02-01). Por fim teremos de introduzir os dias de feriados presentes nos 7 dias anteriores ao dia de previsão, caso não existem deixar em "branco".

No *mainPanel* apenas temos um separador, Previsão Dia, ao contrário do que acontece no separador de Previsão Mensal uma vez que aqui não teremos vários valores para poder

fazer o gráfico. Neste separador temos uma tabela com 2 colunas com a data de previsão e o respectivo valor.

Powered by Luís Costa Importação Base de Dados **Previsão** Estatísticas Sobre

ATENÇÃO! Para poder fazer previsão é necessário fazer primeiro a importação das base de dados.

Dia de Previsão: 2016-02-01

Dia de previsão é feriado?
 Não
 Sim

t-7	t-6	t-5	t-4
443	399	385	373
2016-01-25 t-3	2016-01-26 t-2	2016-01-27 t-1	2016-01-28
356	311	337	
2016-01-29	2016-01-30	2016-01-31	

Dias de feriado nos 7 dias anteriores ao dia de previsão:

(separados por vírgula) Ex: 2016-01-01,2016-02-09,...

Prever

Previsão Dia

Mostrar 10 registos Procurar:

Data	Valor_Previsto
2016-02-01	435.01

Mostrando de 1 até 1 de 1 registos Anterior 1 Seguinte

Recorte Retangular

Figura 29: Separador "Previsão Diária" da Web-App de Previsão do SUHB

6.3.3 Estatísticas

O separador Estatísticas (Figura 30) é diferente dos separadores discutidos anteriormente, pois não têm um *sidebarPanel* mas sim apenas um *mainPanel* dividido em dois separadores:

1. Visualização de Dados;
2. Gráficos Summary.

No primeiro apenas temos um *input* onde inserimos um ficheiro excell com as mesmas variáveis que a base de dados original introduzida no separador Importação Base de Dados cujo *output* será uma tabela com as variáveis: Sexo, Idade, Data Nascimento, Concelho, Data Admissão, Cor Triagem, Destino e Local Residência.

Nesta tabela podes agora filtrar os dados conforme o utilizador desejar, para isso é usado as "caixas" no topo de cada coluna da tabela tal como mostra a Figura 30.

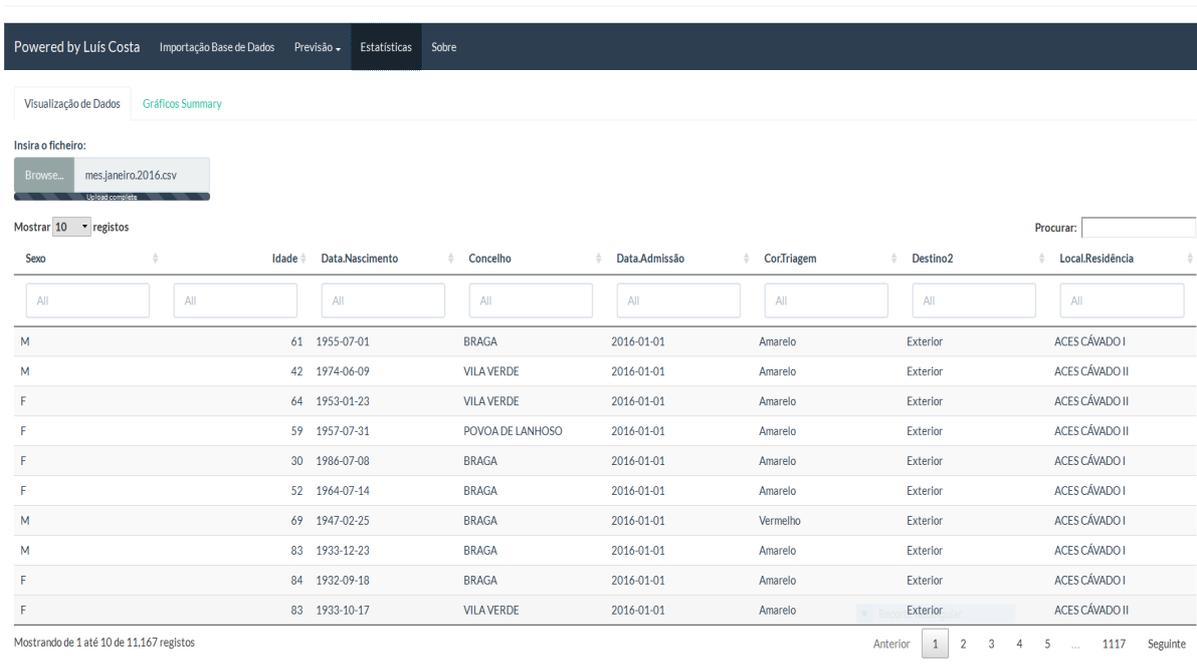


Figura 30: Separador "Estatísticas" da Web-App de Previsão do SUHB

6.3.4 Publicação Aplicação

Várias formas de partilhar uma aplicação em Shiny foram discutidas na Secção 6.2, e uma vez que a aplicação criada para o acompanhamento deste trabalho está finalizada, a forma de partilha da mesma foi feita através do serviço gratuito oferecido pela Shiny (ShinyApps.io).

A aplicação pode ser acedida através do link: <https://luiscosta.shinyapps.io/HospitalFinal/>

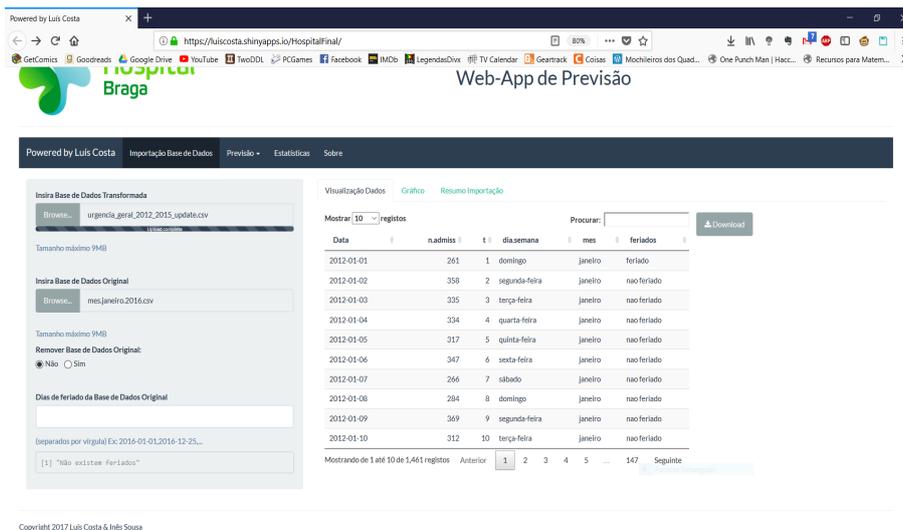


Figura 31: Hospital Braga: Web-App Previsão (publicada em <https://luiscosta.shinyapps.io/HospitalFinal/>)

DISCUSSÃO DE RESULTADOS E CONCLUSÕES

Por último, e depois de muito esforço e noites mal dormidas, chegamos ao final deste trabalho.

Neste trabalho fez-se um estudo sobre a base de dados do SUHB, no qual se pretendia prever o número de admissões diárias no serviço de urgência geral.

A partir da modelação do número de admissões diárias obtemos um primeiro modelo de regressão, no entanto, o termo de erro desse modelo contém erros autocorrelacionados. Uma vez que temos erros autocorrelacionados e estamos a trabalhar com dados de séries temporais, a estimação dos parâmetros do modelo é feita por máxima verosimilhança em que o termo de erro é definido por um processo autoregressivo de ordem 7. Significa que existe correlação semanal, ou seja, o número de admissões de um dia está relacionado com o número de admissões dos sete dias anteriores.

A previsão feita, utilizando este modelo, para o mês de Dezembro de 2016 teve um erro percentual absoluto médio de cerca 8.3%, quando comparamos os valores reais com os valores previstos. Concluimos que o modelo definido tem uma previsão bastante precisa.

Apesar desse ser um bom modelo, neste trabalho tentámos melhorá-lo. Como já referimos anteriormente, o termo de erro do modelo está definido por um processo autoregressivo de ordem 7 e conseqüentemente cada termo deste processo tem a seguinte forma: $\epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-7}$. A solução encontrada foi substituir cada um desses termos por $(Y_{t-1} - \hat{Y}_{t-1}), \dots, (Y_{t-p} - \hat{Y}_{t-p})$, mas ao ser feita esta metodologia o utilizador é obrigado a fazer a previsão diariamente e a fornecer o valor real dos últimos sete dias (anteriores ao dia que pretende prever).

Posto isto, fez-se a previsão para o mesmo período (Dezembro de 2016) que no modelo anterior e obteve-se um erro percentual absoluto médio de 7.8%. Comprova-se uma melhoria na previsão em comparação com o primeiro modelo

Qual modelo utilizar para fazer previsões?

Se a previsão for feita utilizando o primeiro modelo, o utilizador não necessita de introduzir os valores reais observados dos últimos sete dias para poder fazer previsão, no entanto, vai perder precisão na previsão (no caso do mês de Dezembro perdia cerca de 0.5%). Se a previsão for feita utilizando o segundo modelo, o utilizador necessita de intro-

duzir os valores reais observados nos últimos sete dias, mas ganha precisão (no caso do mês de Dezembro de 2016 ganha cerca de 0.5%).

Respondendo à pergunta: Depende da pessoa que estiver a fazer previsão e depende também da precisão que quer obter.

O objectivo deste trabalho foi atingido com sucesso, uma vez que, conseguimos modelar o número de admissões diárias do SUHB. No entanto, os dados deste trabalho são referentes apenas ao período de Janeiro de 2012 até Dezembro de 2016, daí ser necessário estar sempre a actualizar os modelos com os dados mais recentes. Uma forma encontrada para resolver esta situação foi a construção da web-app apresentada no Capítulo 6, onde o utilizador apenas precisa de fornecer a base de dados referente ao período Janeiro de 2017 até ao presente. A aplicação faz a modulação dos dados novamente, obtendo estimativas novas para os parâmetros, de forma a que, a previsão seja o mais correcta possível.

BIBLIOGRAFIA

- JJ Allaire. Application layout guide (shiny). 2014. URL <https://shiny.rstudio.com/articles/layout-guide.html>.
- Chris Beeley. *Web Application Development with R Using Shiny*. Packt Publishing, 2013. URL <https://englianhu.files.wordpress.com/2016/01/web-application-development-with-r-using-shiny.pdf>. 978-1-78328-447-4.
- Bruce L. Bowerman and Richard T. O'Connell. *Forecasting And Time Series: An Applied Approach*. Duxbury Press, 3rd edition, 1993.
- Annette J. Dobson. *An Introduction To Generalized Linear Models*. Chapman Hall/CRC, 2nd edition, 2002. 1-58488-165-8.
- Garrett Grolemond. Shiny html tags glossary. 2017. URL <https://shiny.rstudio.com/articles/tag-glossary.html>.
- James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994. ISBN 0-691-04289-6.
- AM Hopkins J Wojciechowski and RN Upton. Interactive pharmacometric applications using r and the shiny package. 2015. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4394611/>.
- C.D. Lewis. *Industrial and business forecasting methods*. Butterworths, 1982.
- Raquel Menezes. *Sebenta Séries Temporais*. 2015.
- Andrzej Oleś. Interactive data visualization with shiny. 2015. URL <https://www.bioconductor.org/help/course-materials/2015/CSAMA2015/lab/shiny.html>.
- Alan Pankratz. *Forecasting With Dynamic Regression Models*. Wiley-Interscience Publication, 1991. ISBN 0-471-61528-5.
- RStudio. Shiny description, 2017. URL <https://shiny.rstudio.com>.
- Thomas P. Ryan. *Modern Regression Methods*. John Wiley Sons, 1997. 0-471-52912-5.

