UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE FÍSICA
BACHARELADO EM FÍSICA

DANIELLE SCHNEIDER SANTOS

APRENDIZADO DE MÁQUINA: ESTATÍSTICA BAYESIANA EM MÉTODO DE REGRESSÃO LINEAR SIMPLES COM APLICAÇÃO EM MAGNITUDES DE QUASARES

PORTO ALEGRE
2018

DANIELLE SCHNEIDER SANTOS

APRENDIZADO DE MÁQUINA: ESTATÍSTICA BAYESIANA EM MÉTODO DE REGRESSÃO LINEAR SIMPLES COM APLICAÇÃO EM MAGNITUDES DE QUASARES

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Física, pelo Curso de Bacharelado em Física - Pesquisa Básica da Universidade Federal do Rio Grande do Sul - UFRGS.

Orientadora: Profa. Dra. Alejandra Daniela Romero

PORTO ALEGRE

2018

DANIELLE SCHNEIDER SANTOS

APRENDIZADO DE MÁQUINA: ESTATÍSTICA BAYESIANA EM MÉTODO DE REGRESSÃO LINEAR SIMPLES COM APLICAÇÃO EM MAGNITUDES DE QUASARES

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Física, pelo Curso de Bacharelado em Física - Pesquisa Básica da Universidade Federal do Rio Grande do Sul - UFRGS.

Aprovada em:

BANCA EXAMINADORA

Profa. Dra. Alejandra Daniela Romero (Orientadora)
Universidade Federal do Rio Grande do Sul (UFRGS)

Prof. Dr. Jeferson J. Arenzon
Universidade Federal do Rio Grande do Sul (UFRGS)

Prof. Dr. Jose Eduardo da Silveira Costa
Universidade Federal do Rio Grande do Sul (UFRGS)

Às minhas mães: Hebe e Graça.

AGRADECIMENTOS

Agradeço a todos que me apoiaram de alguma forma durante o período da graduação, ajudando-me a estudar, concluir as cadeiras e terminar este trabalho.

Às minhas mães - Graça e Hebe - por serem o alicerce de tudo o que construí, construo e ei de construir. É o incentivo delas que me guia.

Ao Thiago, meu companheiro, por acreditar em mim mais do que eu mesma. Sem ele, eu teria desistido.

Às minhas amigas, em especial à Thayse, Juliana e Mayra. Amizades que construí durante a graduação e que tornaram todos os dias de estudo mais tranquilos e alegres.

À minha orientadora, Alejandra, e ao Rogério, COMGRAD, agradeço pelo suporte em todos os momentos.

Por fim, a todos meus professores e colegas de trabalho que pacientemente entenderam e aceitaram a difícil tarefa de me acompanhar numa jornada, de cinco anos e meio, dividida entre trabalho e faculdade.

“... It’s always darkest before the dawn ...”

(Shake It Out - Florence and the Machine)

RESUMO

Neste trabalho, conceitos de inferência estatística são utilizados para aplicação de métodos de regressão linear simples através de técnicas de aprendizado de máquina. Como principal objetivo, busca-se analisar e comparar o método de regressão linear simples entre duas interpretações diferentes: abordagem estatística bayesiana e abordagem frequentista, também chamada de clássica. Utilizando-se bibliotecas específicas de programação probabilística e aprendizado de máquina para linguagem de programação Python, realiza-se, computacionalmente, a análise para os dois vieses de interesse.

Os dados utilizados para análise são referentes às magnitudes *i* e *z* dos quasares obtidos pelo *Sloan Digital Sky Survey* (SDSS). Estas demonstram forte correlação entre si, sendo possível descrevê-las através de uma reta.

Para a abordagem clássica, espera-se encontrar a linha de regressão que melhor descreva os dados. Para tal, busca-se encontrar uma estimativa única e pontual para os parâmetros de regressão. Em contrapartida, a análise bayesiana prevê que os parâmetros são descritos através de distribuições, ao invés de valores pontuais.

Palavras-chave: Aprendizado de Máquina. Estatística. Estatística Bayesiana. Regressão Linear. Frequentista.

ABSTRACT

In this work, statistical inference concepts are used to apply simple linear regression methods through machine learning techniques. The main objective is to analyze and compare the simple linear regression method between two different interpretations: bayesian statistical approach and frequentist approach, also called classical. Using specific libraries of probabilistic programming and machine learning for the Python programming language, the analysis for the two biases of interest is carried out computationally.

The data used for analysis refer to the magnitudes *i* and *z* of the quasars obtained by the *Sloan Digital Sky Survey*. These show a strong correlation between them, being possible to describe them through a straight line.

For the classical approach, one expects to find the regression line that best describes the data. To achieve this, it is needed to find a unique and punctual estimate for the regression parameters. In contrast, bayesian analysis predicts that the parameters are described through distributions, rather than unique values.

Keywords: Machine Learning. Statistics. Bayesian Statistics. Linear Regression. Frequentist.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Métodos de aprendizado de máquina têm sido utilizados para diversas aplicações com diferentes casos de uso, como sistemas de recomendação, análise de imagens, programação neurolinguística. Um dos principais tipos de tarefas, a regressão, utiliza majoritariamente conceitos estatísticos através da abordagem frequentista. Nesse trabalho, propõe-se demonstrar como incorporar outro viés da estatística, a abordagem bayesiana, ao método de regressão linear simples utilizando conceitos de aprendizado de máquina e programação probabilística.

O problema de regressão proposto utiliza dados do *Sloan Digital Sky Survey* [1] sobre magnitudes dos quasares observados através de diferentes filtros. A análise por regressao linear simples para as duas abordagens é realizada correlacionando as magnitudes observadas através de dois destes filtros.

Os capítulos seguintes versam sobre: Capítulo 2 - referencial teórico utilizado como embasamento para realização da análise proposta, onde encontra-se informações sobre Inteligência Artificial, Inferência Estatística e Métodos de Regressão; Capítulo 3 - metodologia realizada para incorporação do viés frequentista e bayesiano ao problema proposto, composto de: Aquisição de dados, Análise Exploratória e Visualização dos Dados, Regressão Linear frequentista e bayesiana; Capítulo 4 - demonstração dos resultados obtidos; Por fim o Capítulo 5 onde propostas para trabalhos futuros são descritas.

[1] Mais informações em <https://www.sdss.org/>

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Inteligência Artificial e Aprendizado de Máquina

Inteligência Artificial [1], (ou *Artificial Intelligence* (AI)), compreende a ciência e engenharia de criar máquinas capazes de performar ações características à inteligência humana, sendo formadas máquinas inteligentes.

Aprendizado de Máquina, ou *Machine Learning* (ML), por sua vez, é uma área da Inteligência Artificial na qual os sistemas são capazes de aprender e aperfeiçoar tarefas através da experiência, sem serem explicitamente programados. “Diz-se que um programa de computador aprende através da experiência "E" com relação a um conjunto de tarefas "T" e desempenho "P", se seu desempenho em tarefas T, medidos por P, melhoram com a experiência E"(MITCHELL, 1997).

Orientada à utilização, ML é uma aplicação de AI baseada na concepção de que através de dados e informações fornecidas, as máquinas podem aprender por si mesmas. Através da análise dos dados, os algoritmos de aprendizado de máquina melhoram adaptativamente seu desempenho à medida que o número de amostras disponíveis para aprendizado aumenta. Um dos propósitos é fazer com que computadores consigam desempenhar algo natural para seres humanos: Aprender através da experiência.

A parte de treinamento - aprendizado - dos algoritmos é, em sua essência, compreendida pela etapa de inserção de dados para o algoritmo e suas subsequentes iterações de aperfeiçoamento. Nesse processo de aprimoramento, acontece o aprendizado da máquina.

O objetivo fundamental dos algoritmos de aprendizado de máquina é generalizar para além das amostras de dados de treinamento, ou seja, interpretar com sucesso dados desconhecidos anteriormente baseado em informações utilizadas para o aprendizado. De tal forma que, a parte de aprendizado é compreendida pela representação, avaliação e otimização (DOMINGOS, 2012).

2.1.1 Generalizando grupos de aprendizagem

De uma maneira ampla, algoritmos de aprendizado de máquina expressam soluções para tarefas como classificação, regressão e agrupamento (*classification, regression e clustering*). Os algoritmos, por sua vez, são usualmente classificados como algoritmos de aprendizado

[1] Termo cunhado por John McCarthy, apresentado em 1956 em uma conferência em Darthmount (NILSSON, 2009).

supervisionado, não supervisionado e por reforço (*supervised, unsupervised e reinforcement*).

Algoritmos de aprendizado supervisionado utilizam um conjunto de dados mapeados em duplas, entrada e saída, para treinar um modelo a fim de gerar previsões razoáveis de saída para novos dados de entrada. Utiliza-se esse algoritmo quando se tem um conjunto de informações rotuladas disponíveis para treinamento. Dessa maneira, o algoritmo pode aprender como predizer a saída para novos dados de entrada.

Os algoritmos não supervisionados conseguem encontrar padrões intrínsecos nas observações. Utilizado para extrair inferências em conjuntos de informações de entrada desprovidas de saída rotuladas. Algoritmos não supervisionados não utilizam conhecimento algum sobre dados de saída mapeados no processo de predição e classificação, o aprendizado é realizado unicamente através das observações dos dados de entrada (BECKER; PLUMBLEY, 1996).

A ideia do aprendizado por reforço, por sua vez, é de que a máquina aprenda como se comportar em um determinado ambiente executando ações e analisando os resultados obtidos. Desenvolvendo determinados comportamentos através de interações com o ambiente e sendo recompensada ou recebendo penalidades por estes, a máquina é capaz de aprender como se comportar baseando-se na experiência prévia. Nesta modalidade, a máquina está inserida em um ambiente onde é treinada continuamente utilizando a abordagem de tentativa e erro (KAELBLING *et al.*, 1996).

Os algoritmos são, então, classificados de acordo com seu comportamento no qual aprendizado supervisionado é dito guiado a tarefas, o não supervisionado a dados, enquanto o aprendizado por reforço aprende a reagir ao ambiente.

Em relação às soluções, problemas de classificação são compreendidos pela atribuição de classes predefinidas para novas observações com base no aprendizado dos dados prévios. As classes são valores discretos que os novos dados observados podem assumir. Em contrapartida, o agrupamento busca em agregar os dados por encontrar estrutura entre estes antes desconhecidas. De forma que dados com características semelhantes sejam colocados no mesmo grupo, enquanto itens não similares fiquem em grupos distintos.

Já a tarefa de regressão é, basicamente, uma abordagem estatística utilizada para encontrar relações entre variáveis. A modelagem preditiva de regressão busca aproximar uma função de mapeamento (f) das variáveis de entrada (X) para variáveis de saída contínuas (Y). No aprendizado de máquina, assim como em outras áreas, usa-se técnicas de regressão para prever o resultado de um evento, ou seja, realizar inferências a partir de novas observações, com base na

relação entre as variáveis obtidas a partir do conjunto de dados prévios.

2.2 Inferência Estatística

Inferência, na estatística, compreende o processo de formular conclusões a partir de conjuntos de dados. Comumente, tem-se interesse em obter informações sobre determinadas características de um grande grupo de elementos, por exemplo, indivíduos, produtos, preços, elementos, entre outros. O objetivo da inferência estatística é realizar predições sobre estes grupos baseadas em informações referentes a alguns elementos e quantificar o nível de incerteza atrelado a essas predições. No contexto físico, essa situação, como exemplificada em (NEYMAN, 1937), pode ser descrita pelo interesse em saber a quantidade de partículas α emitidas por alguma matéria radioativa. É possível analisar o problema físico através de um modelo matemético, onde a emissão das partículas seja uma função de um parâmetro, como a vida média de um átomo. A inferência, neste caso, reside em utilizar as observações disponíveis para determinar o parâmetro de interesse.

Ao construir modelos estatísticos, considera-se parâmetros e erros de medição como variáveis aleatórias cujas propriedades ou distribuições estatísticas desejamos inferir usando dados medidos. Dessa forma, busca-se entender o comportamento de uma população baseando-se em dados contidos em uma amostra desta. Entende-se por população, todos os elementos ou resultados do grupo de um determinado estudo. A amostra, por sua vez, é qualquer subconjunto contido na população. Assim, assume-se que a população é maior do que o conjunto de dados observados, a amostra. Os dados da amostra são utilizados para desenvolver estimativas das características da população como um todo. De forma que, a Inferência Estatística consiste em fazer afirmações probabilísticas sobre as características do modelo estatístico, que se supõe representar uma população, a partir dos dados de uma amostra aleatória (probabilística) desta mesma população (REIS, 2018), onde as afirmações probabilísticas são associadas a uma probabilidade de ocorrência referente à afirmação.

A inferência estatística consiste, portanto, em selecionar um modelo estatístico que descreve o processo de geração dos dados e, a partir deste, deduzir afirmações. Amplamente, a inferência estatística pode ser compreendida através de dois paradigmas diferentes: frequentista e bayesiano (WASSERMAN, 2004). As inferências freqüentista e bayesiana diferem nos conceitos referentes à natureza das probabilidades, modelos, parâmetros e intervalos de confiança.

A interpretação frequentista, chamada também de clássica, descreve que as probabi-

lidades estão intrinsicamente relacionadas às frequências dos eventos, são definidas com base nas frequências com as quais um evento ocorre ao repetir um determinado experimento por um dado número de vezes. Seguindo essa abordagem, a probabilidade é tida como objetiva e não é atualizada à medida que novos dados são adquiridos. Dessa forma, os parâmetros do modelo estatístico que representa a população são considerados desconhecidos, mas fixos e, portanto, determinísticos. Pela definição descrita no dicionário, probabilidade é o grau de segurança com que se pode esperar a realização de um evento, determinado pela frequência relativa dos eventos do mesmo tipo numa série de tentativas.

Já a perspectiva bayesiana trata as probabilidades como uma distribuição de valores subjetivos que são atualizadas à medida que os dados são observados, ao invés de serem relacionadas a uma frequência como na interpretação frequentista. O conceito de probabilidade é estendido para abranger graus de certeza sobre as afirmações, dado que a probabilidade de um evento é interpretada como a medida do grau de confiança na ocorrência do evento. Os parâmetros populacionais desconhecidos são tratados como variáveis aleatórias com funções densidades associadas e é possível representar informações prévias as observações através de modelos probabilísticos. A inferência bayesiana consiste em combinar informações subjetivas com informações proveniente dos dados observados, através do teorema de Bayes.

2.3 Métodos de Regressão

Utilizada, majoritariamente, para modelos de previsão, a análise de regressão permite estimar as relações entre variáveis e é um dos modos de formulação do modelo estatístico. Portanto, é uma forma de modelagem preditiva que analisa a relação entre variáveis dependentes (Y) e variáveis independentes (X) com o intuito de encontrar vínculos entre estas, através de parâmetros β. Matematicamente (Equação 2.1):

$$Y \sim f(X, \beta) \tag{2.1}$$

A função representada por f pode assumir diferentes formas e depende, basicamente, do número de variáveis independentes, do tipo de variável dependente e da forma de distribuição dos dados a serem expressos pela linha de regressão. Os modelos mais comuns são: regressão linear e regressão logística. A regressão logística é utilizada quando Y é de natureza binária ou dicotômica, X de natureza categórica ou não categórica. Já na regressão linear, a variável

dependente é contínua, a variável independente pode ser contínua ou discreta e a linha de regressão é linear.

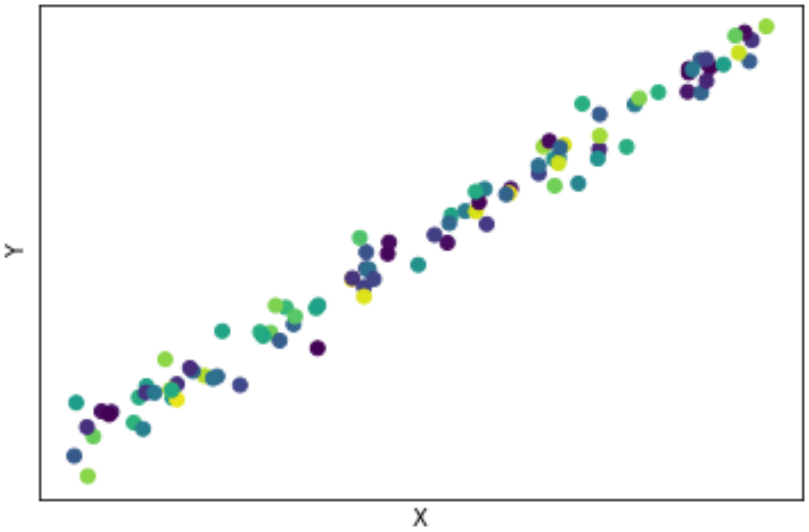
2.3.1 Regressão Linear Simples - Abordagem Frequentista

Na abordagem clássica do método de regressão linear simples, o modelo usado para descrever a população - o modelo estatístico - é representado por uma combinação linear dos parâmetros - coeficientes de regressão - na forma da Equação 2.2. Esta, descreve a relação entre as variáveis explicativas e resposta, X e Y respectivamente.

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad (i = 1, 2, ..., n) \tag{2.2}$$

Em uma representação gráfica, para que os dados sejam representados por uma regressão linear simples, espera-se que os diagramas de dispersão sejam da forma disposta na Figura 1. Uma correlação linear entre os dados.

Figura 1 – Correlação entre os dados (X e Y) representada através de um gráfico de dispersão.



Na Equação 2.2, Y_i é a variável dependente, X_i é a variável resposta, e_i é relativo aos fatores residuais e erros, n indica o tamanho da amostra, e β_0 e β_1 são os parâmetros a serem estimados. O coeficiente β_0 equivale ao ponto onde ocorre a interceptação da reta com o

eixo vertical, coeficiente linear, e é dito constante da equação de regressão. Já β_1 representa a declividade, coeficiente angular, da reta e é chamado coeficiente de regressão.

Estabelecido o modelo estatístico, busca-se obter a linha de regressão que melhor descreve a população. Para isso, procura-se estimar os valores dos parâmetros da equação linear que ajustam o modelo da melhor maneira possível. Dessa forma, β_0 e β_1 são obtidos utilizando técnicas de estimativa de parâmetros.

O propósito da estimativa de parâmetros é descrever algum aspecto desconhecido da população. Dado um modelo estatístico representado por uma determinada família de funções, busca-se encontrar os parâmetros necessários a família de funções que melhor descrevem o conjunto de dados observados. Para tal, retira-se uma amostra aleatória desta população e, utilizando-se técnicas de estimativas de parâmetros, procura-se obter uma estimativa do parâmetro de interesse associada a uma probabilidade de qua a estimativa esteja correta.

Sendo o modelo estatístico da forma linear da Equação 2.2, b_1 e b_0 representam os estimadores dos coeficientes β_1 e β_0 respectivamente. Estes, serão obtidos a partir de uma amostra de tamanho n da população e definem uma estimativa geral do modelo escolhido através da Equação 2.3.

$$\hat{Y} = b_0 + b_1 X \tag{2.3}$$

O método mais utilizado para estimar os coeficientes angular e linear da reta no modelo de regressão linear simples é o método de mínimos quadrados (MMQ). Neste, busca-se construir a estimativa geral para a reta de regressão através da minimização da soma dos quadrados da diferença entre os valores reais da amostra Y_i e os correspondentes sobre a reta estimada $\hat{Y}_i$. Ou seja, consiste em minimizar a soma dos quadrados dos resíduos: $\sum_{i=1}^{n} e_i^2$. Assumindo os dados como sendo homocedásticos, esse objetivo é representado tal qual a Equação 2.4.

$$S(b_0, b_1) = \sum_{i=i}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{2.4}$$

Para minimizar a Equação 2.4, iguala-se as derivadas de $S(b_0, b_1)$ em relação a b_0 e b_1 a zero, Equações 2.5, 2.6.

$$\frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \tag{2.5}$$

$$\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \tag{2.6}$$

Considerando a utilização da representação matricial dos dados, onde o modelo é descrito matematicamente pela Equação 2.7:

$$Y = X\beta + e \tag{2.7}$$

sendo

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \text{e} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{2.8}$$

Reescrevendo a Equação 2.4 para representação matricial:

$$S(\beta) = (y - X\hat{\beta})^T (y - X\hat{\beta}) \tag{2.9}$$

onde $\hat{\beta}$ assume o papel de estimadores de β.

Diferenciando a Equação 2.9 com respeito a β e igualando a zero:

$$\frac{\partial S}{\partial \hat{\beta}} = -2X^T(y - X\hat{\beta}) = 0 \tag{2.10}$$

Obtem-se, por fim:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.11}$$

onde cada estimador assume um único valor.

Dessa forma, a inferência sobre novos dados se dará a partir da Equação 2.12

$$\hat{Y} = \hat{\beta}^T X \tag{2.12}$$

2.3.2 Regressão Linear Simples - Abordagem Bayesiana

Na abordagem Bayesiana, a regressão linear é formulada através de distribuições de probabilidade ao invés de estimativas pontuais como a abordagem clássica. A variável resposta

Y não é estimada como um único valor, mas assume-se que as respostas são amostradas a partir de uma distribuição de probabilidade como a Equação 2.13.

$$Y \sim N(\beta^T X, \sigma^2 I) \tag{2.13}$$

A variável dependente Y, é gerada a partir de uma distribuição gaussiana (normal) caracterizada pela média e variância. A média é o produto entre os parâmetros β e variáveis independentes X, enquanto a variância é o quadrado do desvio padrão σ. Neste modelo, assim como a variável resposta é assumida como uma amostra de uma distribuição, os parâmetros também o são. Ou seja, no viés bayesiano, os dados não observados e os parâmetros desconhecidos do modelo são tratados de maneira semelhante.

Para encontrar as distribuições dos parâmetros do modelo, a inferência bayesiana utiliza o Teorema de Bayes para combinar informações prévias ao experimento e dados de amostra com o intuito de deduzir propriedades e conclusões sobre um parâmetro de interesse a partir de dados de entrada, X, e saída, Y.

Teorema 2.3.1 (Bayes) *A probabilidade de qualquer evento é a razão entre o valor em que uma expectativa, dependendo do acontecimento do evento, deve ser computada, e o valor do que se espera que aconteça.*

Matematicamente, Laplace[2] formulou como:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.14}$$

Onde, no contexto desse trabalho:

$$P(\beta|y,X) = \frac{P(y|\beta,X)P(\beta|X)}{P(y|X)} \tag{2.15}$$

O lado esquerdo da Equação 2.15 é chamado de probabilidade *a posteriori* ($P(\beta|y,X)$), os termos do numerador são ditos probabilidade *a priori* ($P(\beta|X)$) e verossimilhança ($P(y|\beta,X)$).

A probabilidade *a priori* reflete o grau de certeza, crença, e por consequência a incerteza, de β inicialmente, antes de realizar o experimento. Qualquer informação que se tenha inicialmente sobre o parâmetro é tratada como a probabilidade *a priori*, por exemplo pode-se considerar informações baseadas em modelos anteriores com dados semelhantes. Quando não

[2] Pierre-Simon Laplace (1749–1827), contribuiu com trabalhos nas áreas de matemática, estatística, física e astronomia.

se possui qualquer informação referente aos parâmetros antes da realização do experimento, visualização dos dados, costuma-se utilizar uma função de distribuição de probabilidade não informativa, como a distribuição uniforme. A habilidade de incorporar crenças prévias é uma das principais diferenças entre a inferência bayesiana e frequentista.

A verossimilhança é interpretada da mesma forma que no viés clássico, podendo ser também conhecida como evidência. Diz respeito a distribuição característica aos dados observados.

O denominador da Equação 2.15, como não depende dos parâmetros do modelo, é tratado como uma constante de normalização para a probabilidade *a posteriori* e reflete a probabilidade com qual pode-se obter qualquer dado.Explicitamente consiste em integrar o lado direito da Equação 2.15 através de todos os valores possíveis dos parâmetros, tal qual a Equação 2.16.

$$P(y|X) = \int_{\beta} P(y|\beta, X) \times P(\beta|X) d\beta \tag{2.16}$$

Substituindo 2.16 em 2.15, obtem-se:

$$P(\beta|y, X) = \frac{P(y|\beta, X)P(\beta|X)}{\int_{\beta} P(y|\beta, X) \times P(\beta|X) d\beta} \tag{2.17}$$

A solução da estimativa de parâmetros na abordagem Bayesiana é compreendida pela probabilidade *a posteriori* que quantifica a incerteza e expressa a distribuição de probabilidade dos parâmetros do modelo tendo sido observados os dados, X e Y.

Uma das formas de obter a distribuição *a posteriori* sem calcular analiticamente a integral do denominador é utilizando métodos de Monte Carlo via cadeias de Markov (MCMC)[3]. MCMC é um método de amostragem computacional que permite caracterizar uma distribuição sem conhecer todas as propriedades matemáticas desta ao amostrar aleatoriamente valores da distribuição.

[3] Mais informações sobre Monte Carlo via cadeia de Markov podem ser encontradas na literatura, como em (NEWMAN; BARKEMA, 1999) não sendo o objetivo deste trabalho.

3 METODOLOGIA

O processo utilizado, visando a aplicação da estatística bayesiana em um problema de regressão linear simples, envolve as seguintes etapas: Aquisição, análise e processamento dos dados; Regressão Linear - abordagem clássica; Regressão Linear - abordagem bayesiana; e Análise dos resultados. Para realizar cada etapa deste processo, diferentes ferramentas foram utilizadas.

A liguagem de programação Python[1] foi usada como base para implementação em conjunto com o *Jupyter Notebook*[2], um ambiente computacional interativo capaz de mesclar execução de código, equações, texto, executar simulação numérica, modelagem estatística, aprendizado de máquina e suportar plotagem de gráficos, visualização e análise de dados. Para realizar a análise dos dados, foram utilizadas as bibliotecas: *matplotlib* e *seaborn* como ferramentas na geração dos gráficos; e *pandas*, para leitura dos dados e métricas.

A regressão linear - abordagem clássica - foi realizada através da função de regressão linear da biblioteca *scikit-learn*, que compreende vários algoritmos de aprendizado de máquina em código aberto para a linguagem de programação Python. Já a regressão linear - abordagem bayesiana - foi realizada utilizando a biblioteca *pymc3*, um pacote do ecossistema Python especializado para modelagem estatística bayesiana e aprendizado de máquina probabilístico, focando em Monte Carlo via cadeia de Markov e algoritmos de inferência variacional.

3.1 Aquisição dos Dados

O *Sloan Digital Sky Survey* (SDSS, <https://www.sdss.org/>) é um dos maiores, mais detalhados e mais citados levantamentos astronômicos que já existiu, com o objetivo de expandir nossa compreensão sobre a evolução e estrutura em larga escala do Universo, a formação de estrelas e galáxias, a história da Via Láctea e a ciência por trás da energia escura" (FOUNDATION, 2018).

O sistema de fotometria[3] utilizado pelo SDSS compreende cinco filtros passa-faixa (*u, g, r, i, z*), com bandas passantes não sobrepostas entre 3000Å a 11000Å. O filtro *u* tem um pico em 3551Å; *g* compreende uma banda verde-azul centrada em 4686Å; *r* é passa-banda

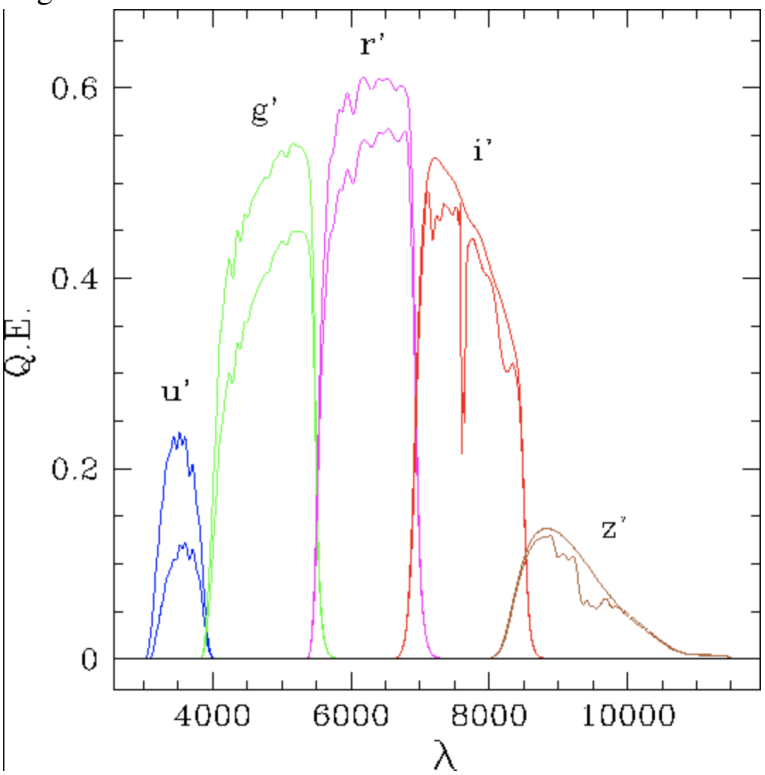
[1] Disponível em: https://www.python.org/

[2] Mais informações em: http://jupyter.org/

[3] A palavra composta dos afixos gregos foto- (“luz”) e –metria (“medida”), representa uma técnica da astronomia relacionada à medição do fluxo ou intensidade da radiação eletromagnética de um objeto astronômico (STERKEN; MANFROID, 1992)

vermelho centrado em 6166Å; *i* é um filtro do extremo vermelho centrado em 7480Å; por fim, o filtro *z* é referente ao infravermelho próximo com pico em 8932Å (FUKUGITA *et al.*, 1996). As curvas características de cada um dos filtros estão ilustradas na Figura 2.

Figura 2 – Curvas características dos filtros do SDSS.



Fonte: https://arxiv.org/pdf/astro-ph/9809085.pdf

Entre os objetos astronômicos observados pelo *Sloan Digital Sky Survey*, encontram-se os quasares - *quasi-stellar radio object* - objetos quase estelares com emissão em rádio. Quasares são um tipo de galáxia ativa, sendo os objetos mais luminosos do universo. Descobertos, como intensas fontes de rádio, com aparência ótica aproximadamente estelar, azuladas. São objetos extremamente compactos e luminosos, emitindo mais energia do que centena de galáxias juntas. São forte fontes de rádio, variáveis, e seus espectros apresentam um notável deslocamento para o vermelho devido ao efeito da expansão do Universo indicando que eles estão se afastando a velocidades muito altas, de até alguns décimos da velocidade da luz. O espectro dos quasares, e das galáxias ativas em geral apresentam linhas largas e estreitas devidas ao efeito Doppler [4] que indicam que o material na fonte está se movendo a velocidades de $\sim$ 5000 km/s e $\sim$ 500 km/s, respectivamente. No modelo mais aceito para os quasares, eles são galáxias com um buraco negro supermasivo no centro que acreta gás e estrelas da sua vizinhança. O processo de

[4] Para um corpo luminoso aproximando-se ou afastando-se do observador, o comprimento de onda da luz diminui ou aumenta, respectivamente em relação àquele observado em laboratório.

acreção da lugar à emissão de alta energia e enquanto a matéria acelera, espiralando no disco de acreção. Por ação do campo magnético, parte da matéria é ejetada em forma de jatos a velocidades relativísticas (KEPLER; SARAIVA, 2014).

A análise de regressão, compreendida neste trabalho, é realizada com dados referentes a 1000 quasares disponíveis pelo SDSS (através do *SkyServer's SQL Search tool*), com informações relativas às magnitudes observadas em cada um dos cinco filtros, assim como o valor referente ao *redshift*[5] respectivo a cada quasar.

3.2 Análise Exploratória e Visualização dos Dados

A etapa de Análise Exploratória compreende a prática de descrever os dados estatisticamente e através de técnicas de visualização, com o objetivo de ressaltar os pontos mais importantes do conjunto de dados para uma análise posterior. Este processo considera a análise do dados disponíveis por diferentes aspectos, a descrição e o condensamento das informações sem fazer nenhum tipo de julgamento referente ao conteúdo dos dados.

Esta análise é um passo importante a ser dado antes de se começar o tratamento de um conjunto de dados com técnicas de ML ou o modelagem estatística, já que características mais evidentes aos dados são notadas e descritas neste processo. Basicamente, a análise consiste em obter confiança no *dataset* a fim de ter conhecimento da sua estrutura.

Os dados do SDSS são obtidos em formato texto, tendo a separação entre chave e valor sendo feita por vírgulas, ou seja, um arquivo CSV - *Comma Separated Values*, onde as incertezas relativas às medidas são desprezadas. Utilizando a função "*read_csv*"da biblioteca *pandas*, os dados podem ser lidos e visualizados de acordo com a Tabela 3, que demonstra cinco dos 1000 dados disponíveis.

Figura 3 – Dados referentes aos quasares obtidos pelo SDSS, contendo informações sobre id característico a cada quasar, magnitudes u, g, r, i, z e redshift.

	objid	modelmag_u	modelmag_g	modelmag_r	modelmag_i	modelmag_z	z
995	588848900452843577	19.167	19.054	18.963	19.004	18.906	0.951
996	587722983890878544	20.860	20.422	20.317	19.964	19.725	0.487
997	587722984427683972	20.431	20.312	20.243	19.981	19.900	1.642
998	588848900989649099	22.017	20.145	19.965	19.575	19.387	2.884
999	587722983354138798	20.583	20.209	20.094	20.095	19.802	0.698

[5] Desvio para o vermelho. No caso dos quasares, este é relacionado à expansão do universo, através do efeito Doppler.

Uma forma de análise exploratória de dados é através da matriz de correlação, que consiste em uma tabela que exibe a correlação entre conjuntos de variáveis. Cada variável randômica (X_i) nesta tabela é correlacionada com cada um dos outros valores na tabela (X_j). Através do processo de criação e análise da matriz de correlação, pode-se observar quais pares de variáveis tem uma correlação mais forte, que é usada para investigar a dependência entre múltiplas variáveis ao mesmo tempo.

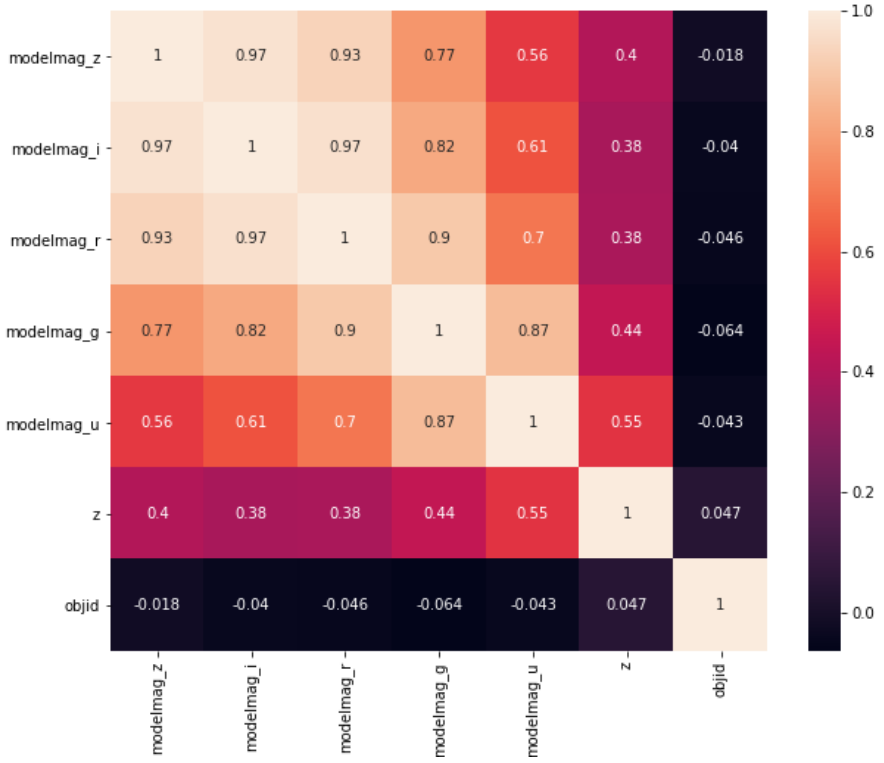
Utilizando a biblioteca *pandas* pode-se calcular a matriz de correlação por três diferentes métodos: Pearson, Kendall e Spearman[6]. Utiliza-se o método de pearson, onde os coeficientes de correlação são calculados com base na dependência linear entre duas variáveis. Em outros termos, estes coeficientes quantificam o nível ao qual a dependência entre duas variáveis pode ser descrita por uma linha. Matematicamente (Equação 3.1):

$$\rho_{X,Y} = \frac{\sum(X_i - \overline{X})\sum(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \tag{3.1}$$

Com as bibliotecas *seaborn* e *matplotlib*, é possível visualizar a matriz de correlação calculada, como é mostrado na Figura 4

[6] Os métodos de correlação de Kendall e Spearman são utilizados para problemas não-paramétricos

Figura 4 – Matriz de correlação entre variáveis de magnitude e redshift dos quasares.



O coeficiente de correlação entre um par de variáveis pelo método de Pearson é denominado ρ e pode variar entre 1 e -1. Quanto mais próximo de 1 for ρ, mais um aumento em uma variável se associa a um aumento na outra. Por outro lado, quanto mais próximo de -1 for ρ, mais o aumento em uma variável resultaria em diminuição na outra. Portanto, o sinal de ρ corresponde à direção da correlação, enquanto o valor corresponde à intensidade da correlação. Uma relação perfeitamente linear (ρ = -1 ou 1) significa que uma das variáveis pode ser perfeitamente representada por uma função linear da outra variável.

Para que os dados possam ser descritos através de uma regressão linear simples da melhor maneira possível, tem-se interesse em dados com alto grau de correlação entre si, visto que o modelo estatítico é representado por uma combinação linear dos parâmetros. Pela análise da matriz de correlação, vê-se que as magnitudes *i*, *r* e *z* apresentam forte linearidade entre si. Os gráficos de dispersão (Figuras 6, 5 e 7) demonstram as correlações lineares correspondentes.

Para análise, escolhe-se a comparação das magnitudes i e z dos quasares, como mostrado na Figura 6, assim como feito no estudo "Modern Statistical Methods for Astronomy" (FEIGELSON; BABU, 2012), onde é demonstrado que, como esperado, existe uma correlação

entre os brilhos de bandas espectrais adjacentes.

Figura 5 – Gráfico de dispersão entre as magnitudes z e r.

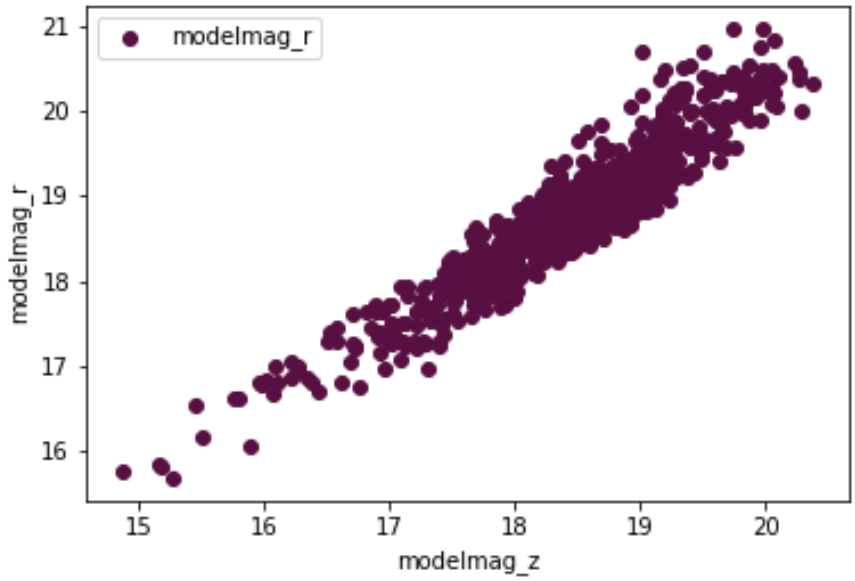


Figura 6 – Gráfico de dispersão entre as magnitudes i e z.

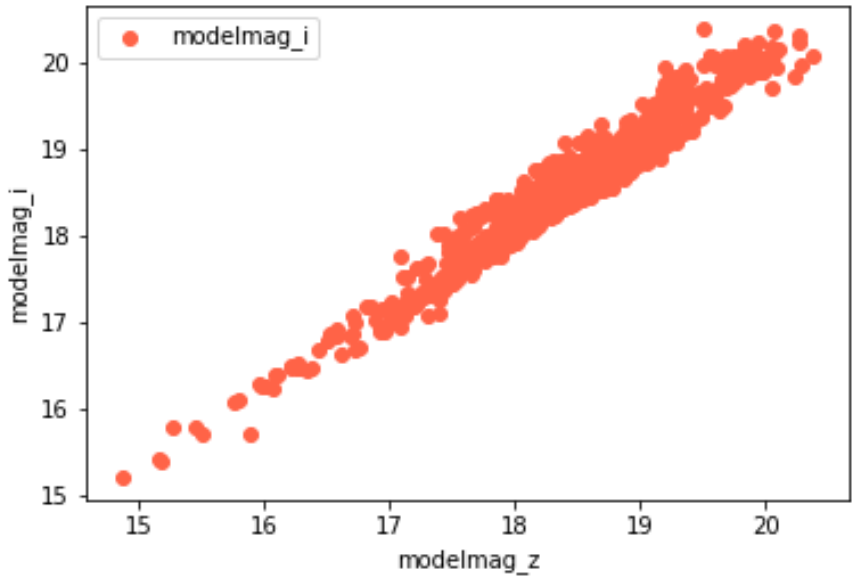
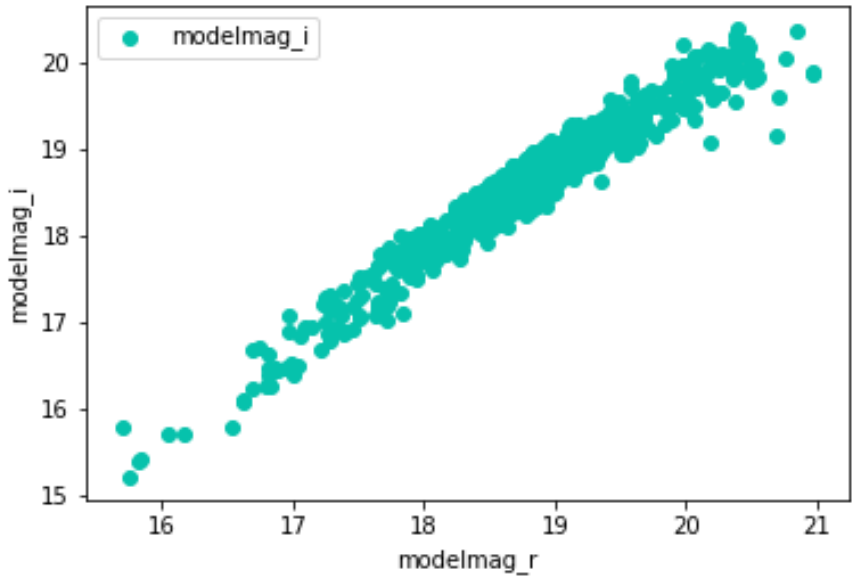


Figura 7 – Gráfico de dispersão entre as magnitudes r e i.



3.3 Regressão Linear Frequentista

A biblioteca *sckitlearn* possibilita a aplicação do método de regressão linear utilizando o estimador pelo método dos mínimos quadrados. Primeiro, divide-se os dados em uma proporção de 80% para treinamento e 20% para validação. Isto é feito para que se possa avaliar o quão bem o modelo performa a partir de dados não vistos durante o treinamento.

A parte considerada para treinamento é utilizada para encontrar os parâmetros que caracterizam a melhor linha de regressão que ajusta e descreve os dados. Obtêm-se, então, os coeficientes de regressão β e β_0 respectivos ao coeficiente angular e interceptação do eixo vertical da reta, calculados pelo método dos mínimos quadrados (Equações 2.4, 2.5, 2.6).

Chama-se a função *score* para obter a acurácia do modelo, uma medida do quão bem os resultados observados são replicados pelo modelo, e é dada pela proporção de variação de saídas total modeladas com sucesso. A acurácia é calculada pelo coeficiente de determinação, R^2 que indica o quanto o modelo estatístico escolhido é capaz de explicar os dados coletados, através da medida da proporção de variabilidade em Y que é explicada por X.

O coeficiente R^2 é definido como $(1 - \frac{u}{v})$, onde u é a soma dos quadrados dos resíduos e v é a soma total dos quadrados, como definido em (3.2) e (3.3) respectivamente. O melhor resultado possível é 1, podendo ser negativo já que o modelo pode ser arbitrariamente pior. Um modelo constante que sempre prediz o valor esperado de Y independente das características

da entrada, teria o valor de R^2 igual a 0.

$$u = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.2}$$

$$v = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.3}$$

O Erro Quadrático Médio (MSE) de um modelo representa uma medida do quão perto uma reta está de representar os dados. Na biblioteca *sckitlearn*, este erro é calculado a partir da função *mean_squared_error* utilizando os 20% dos dados destinados para teste. Utiliza-se o modelo escolhido para inferir valores para a magnitude *i* a partir de valores de magnitude *z*. A diferença entre o valor estimado e o valor real para todos os pontos disponíveis (nos 20% reservados para validação) é utilizada para calcular o erro. Portanto, quanto menor o MSE, mais perto se está de achar a reta que melhor representa os dados.

3.4 Regressão Linear Bayesiana

O princípio do processo de regressão linear bayesiana está em realizar um mapeamento dos dados de treinamento. Nesta etapa, os dados de saída e de entrada são associados especificando que a magnitude i (*modelmag_i*) é uma função da magnitude z (*modelmag_z*), tal qual a Equação (3.4).

$$modelmag\_i \sim modelmag\_z \tag{3.4}$$

O próximo passo, é definir a distribuicão *a priori* dos parâmetros do modelo e interpretar os dados disponíveis para treinamento - as observações das magnitudes dos quasares - como distribuição de verossimilhança. Estas são modeladas através de distribuições normais, gaussianas.

A biblioteca *pymc3* (J. *et al.*, 2016) possibilita a construção de um modelo linear generalizado a partir da fórmula que associa os dados de entrada e saída. Neste modelo são adicionadas variáveis aleatórias para os coeficientes angular e linear de regressão, assim como para a variância.

Com o modelo generalizado construído, para realizar a inferência dos parâmetros do modelo de regressão linear simples no viés bayesiano, utiliza-se métodos de Monte Carlo

via Cadeia de Markov para obter amostras sobre a distribuição posterior. Este método envolve a geração de uma sequência de parâmetros, amostrados conforme uma distribuição, a partir de um valor de parâmetro de início.

O método específico utilizado para performar a amostragem através do método de Markov - Monte Carlo (*Markov chain Monte Carlo*, MCMC) é escolhido automaticamente pela biblioteca, entretanto são especificados o número de amostras de interesse e o número de cadeias de Markov utilizadas: 2000 e 2 repectivamente. Geradas as amostras, plota-se as distribuições *a posteriori* de cada parâmetro pelas funções *traceplot* e *plot_posterior*.

Como última etapa, faz-se uma análise das retas de regressão obtidas. Ao contrário da análise frequentista, na qual busca-se uma única reta de regressão, na abordagem bayesiana uma plotagem preditiva faz uso das amostras obtidas da distribuição a posteriori e traça uma linha de regressão para cada conjunto de valores possíveis. A distribuição das linhas demostra a incerteza dos parâmetros do modelo: quanto mais espalhadas as retas plotadas, maior a incerteza do modelo escolhido naquele ponto.

4 RESULTADOS

Os resultados descritos nesta seção são referentes à regressão linear simples para magnitudes de quasares utilizando as abordagens frequentista e bayesiana.

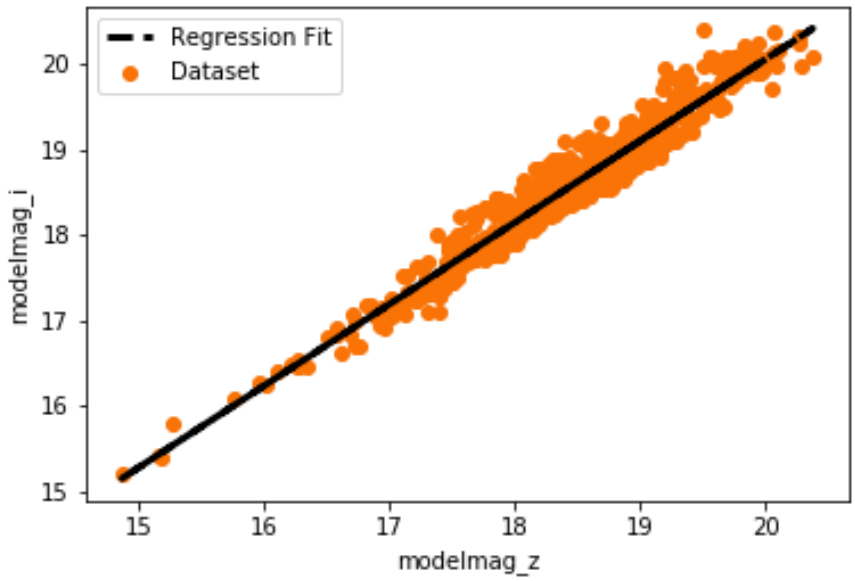
4.1 Resultados - Abordagem Frequentista

No viés clássico, a linha de regressão que melhor representa os dados ocorre quando o coeficiente angular (β) assume o valor de 0.95931138 e o coeficiente linear (β_0) assume o valor 0.88099638. Desta forma, a população é representada pelo modelo estatístico como a Equação 4.1

$$modelmag\_i = (0.95931138) modelmag\_z + 0.88099638 \tag{4.1}$$

Graficamente, o modelo que exprime os dados é representado pela reta mostrada na Figura 8.

Figura 8 – Dados das magnitudes i e z representados através de um diagrama de dispersão e a Linha de regressão correspondente.



A análise de acurácia demonstra que no modelo escolhido para representar a população, o coeficiente R^2 assume um valor de 0.946058559726. Portanto, por volta de 95% da variabilidade em Y pode ser explicada por X. Pelo erro médio quadrático obtido observa-se que

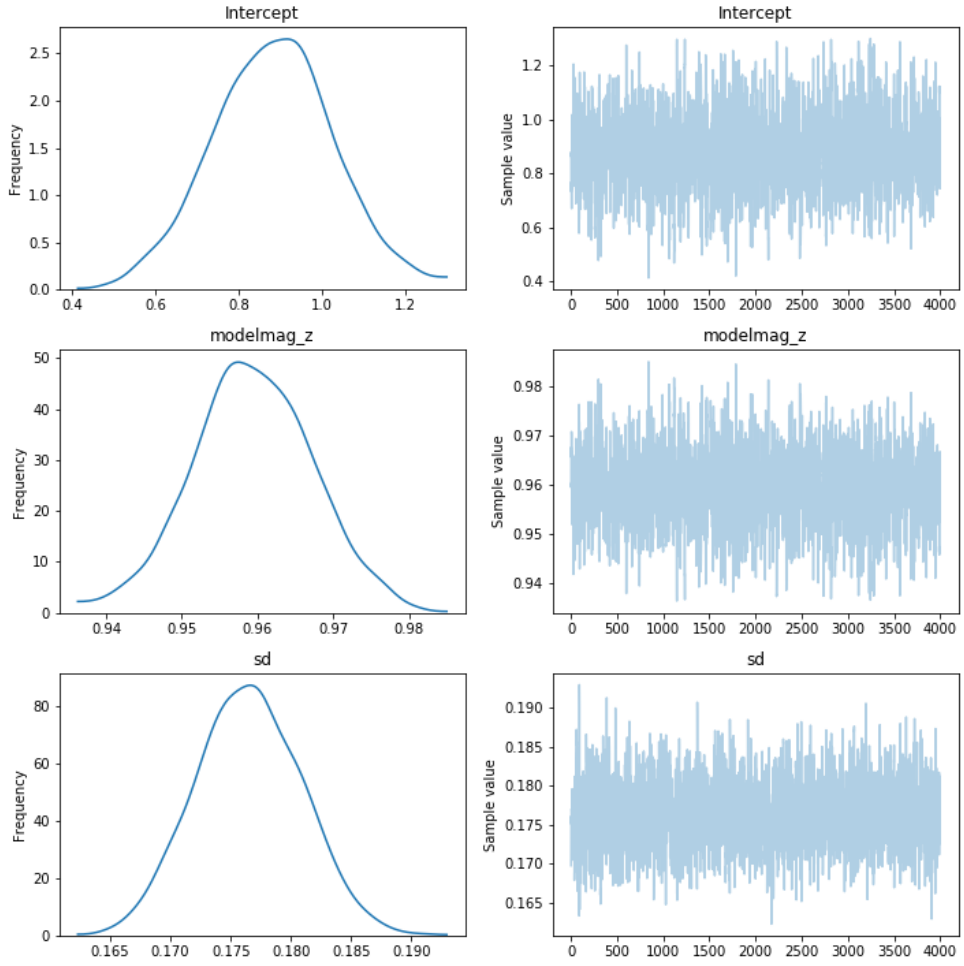
o modelo prediz, em média, um valor para magnitude i que difere de 0.16721311008477635 do valor esperado.

De acordo com o modelo estatístico escolhido para representar a população, a obtenção da magnitude *i* através de inferência realizada com base em um valor 18.014 de magnitude *z*, utilizando dados não vistos pelo modelo na parte de treinamento, resulta no valor 18.162, que difere de 0.148 do valor esperado. Essa diferença está dentro do previsto como desvio padrão: 0.16.

4.2 Resultados - Abordagem Bayesiana

Ao contrário do que acontece na abordagem frequentista, a regressão linear bayesiana não estima um valor único para cada parâmetro. Nesta análise, cada parâmetro do modelo é associado a uma distribuição de probabilidade. A Figura 9 mostra as distribuições inferidas para cada parâmetro utilizando a inferência bayesiana.

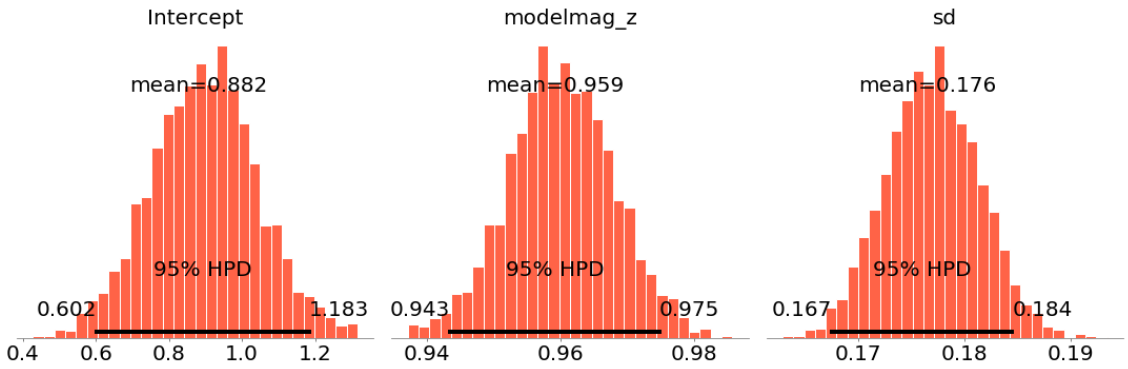
Figura 9 – Distribuições associadas ao coeficiente linear(Intercept), ao coeficiente angular que acompanha a magnitude z(modelmag_z) e ao desvio padrão.



O lado esquerdo da Figura 9 consiste nas distribuições marginais para cada parâmetro de interesse. Para o coeficiente linear, β_0, nota-se que o máximo da distribuição está em torno de 0.9, próximo ao encontrado na abordagem clássica. A estimativa para o coeficiente angular, tem uma distribuição com um máximo em 0.96. Por fim, é disposto o desvio padrão com máximo em 0.176. Já o lado direito, refere-se aos gráficos dos vetores de amostra produzidos pelo procedimento de amostragem do método de Monte Carlo via cadeias de Markov.

Outra maneira de visualização, é através de gráficos de histograma, como os demonstrados na Figura 10.

Figura 10 – Histograma das distribuições associadas ao coeficiente linear(Intercept), ao coeficiente angular que acompanha a magnitude z(modelmag_z) e ao desvio padrão(sd).

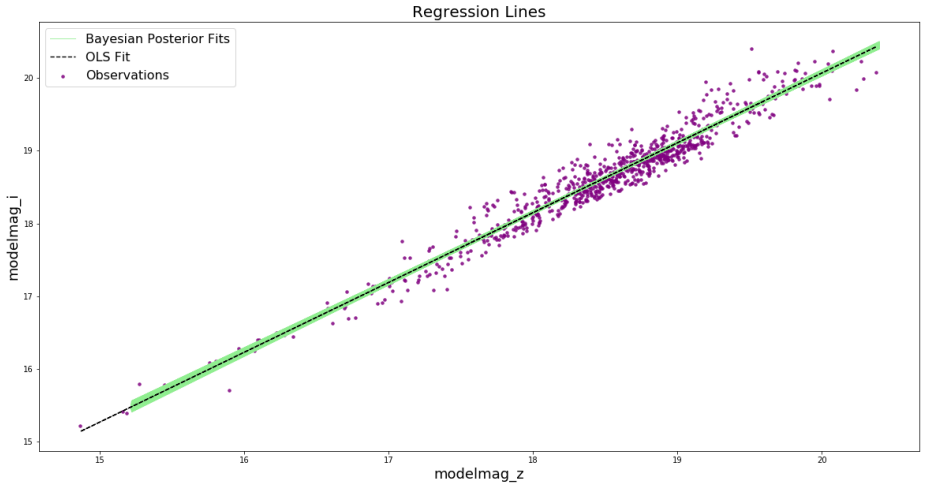


Em todos os parâmetros há uma variância associada a cada distribuição, expressando a existência de um grau de incerteza relativo a cada um dos valores.

Diferentemente do viés frequentista, onde busca-se encontrar uma única linha de regressão, sendo esta a melhor para representar os dados, na regressão linear bayesiana, são descritas várias linhas, cada uma representando uma estimativa diferente dos parâmetros do modelo. A variabilidade das linhas representa a incerteza das estimativas.

Na Figura 11, são demonstradas as linhas inferidas para a regressão linear simples para as magnitudes i e z dos quasares na abordagem bayesiana, a linha de regressão obtida na abordagem frequentista e os dados utilizados.

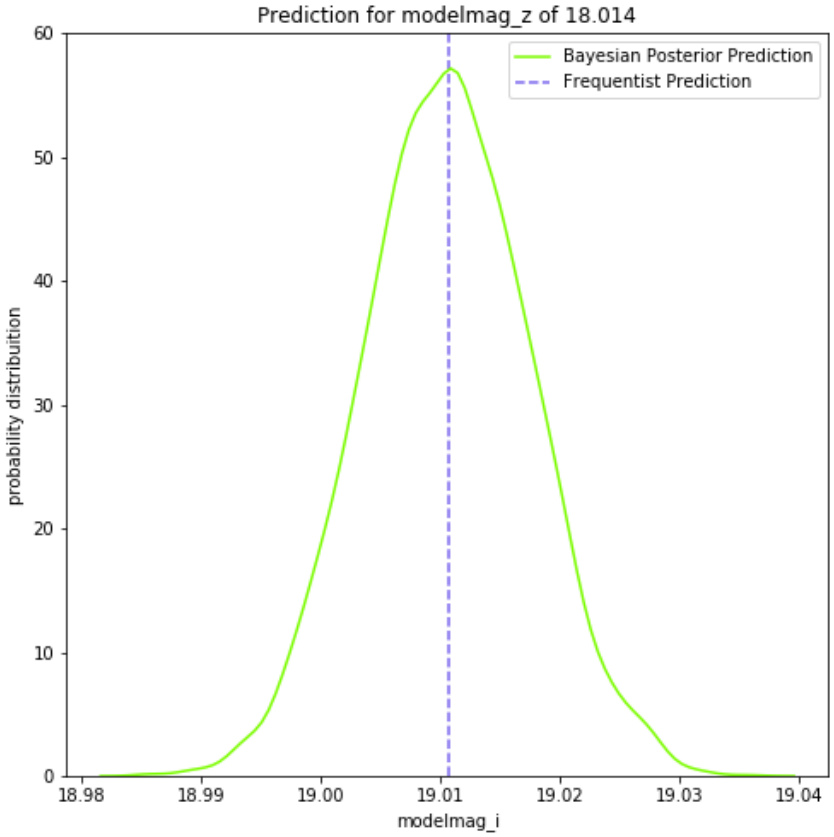
Figura 11 – Linhas de regressão linear simples para a abordagem bayesiana (verde) e frequentista (preto).



Utilizando o modelo estatístico escolhido para realizar uma predição para a mag-

nitude *i* quando um valor de 18.014 para magnitude *z* é observado, obtem-se a distribuição caracterizada na Figura 12, que tem um máximo em 18.162.

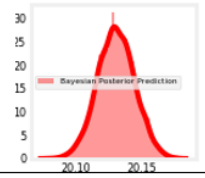
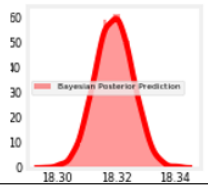
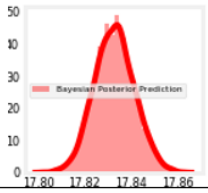
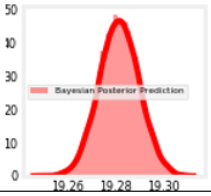
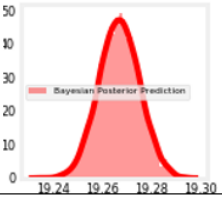
Figura 12 – Comparação entre a predição da distribuição para magnitude i dado o valor de 18.014 para magnitude z com a predição realizada pela abordagem frequentista.



4.3 Comparação entre abordagens

A comparação entre as abordagens é expressada pelas Figuras 13, 14. A tabela compreendida na Figura 13, demonstra os resultados obtidos quando realizadas inferências sob a magnitude *i* através das duas abordagens.

Figura 13 – Dados obtidos ao realizar inferências para a magnitude i com base em valores de magnitude z.

Magnitude z	Magnitude i	Magnitude i Frequentista	Distriuição magnitude i Bayesiana	Magnitude i média bayesiana	Diferença frequentista	Diferença bayesiana
20.066	20.133	20.13053853		20.13038195	0.000246147	0.00161805
18.178	18.408	18.31935865		18.31931111	0.08864135	0.08868889
17.671	17.734	17.83298778		17.83296954	0.09898778	0.09896954
19.181	19.179	19.28154796		19.28144250	0.10254796	0.1024425
19.166	19.36	19.26715829		19.26705370	0.09284171	0.0929463

A Figura 14, por sua vez, expressa os diferentes valores obtidos para os coeficientes de interesse ao modelo estatístico. Como a abordagem bayesiana infere uma distribuição ao invés de um valor pontual, a média da distribuição foi utilizada para comparação. Dessa forma, obtêm-se uma diferença de 0.001003, 0.000057 e 0.009235 entre as abordagens frequentista e bayesiana para o coeficientes lineares, angulares e desvio padrão respectivamente.

Figura 14 – Comparação entre os valores obtidos para os coeficientes linear, angular e desvio padrão utilizando as duas abordagens de interesse.

	Coeficiente Angular	Coeficiente Linear	Desvio Padrão
Abordagem Frequentista	0.959311	0.880996	0.167213
Abordagem Bayesiana	0.959254	0.881999	0.176448

5 CONCLUSÕES

Neste trabalho, aplicou-se conceitos de inferência estatística e aprendizado de máquina para modelar a correlação entre as magitudes dos quasares relativas a dois filtros. Descreveu-se o problema proposto através de uma regressão linear simples, no qual supõe-se que a magnitude i é uma função da magnitude z. Utilizando técnicas de aprendizado de máquina, os parâmetros do modelo estatístico para descrever os dados formam encontrados utilizando abordagens estatística distintas: frequentista e bayesiana.

As vantagens em utilizar a abordagem bayesiana residem em ser possível incorporar informações prévias às observações, atribuir distribuições aos parâmetros de interesse ao invés de uma estimativa pontual, o que torna possível atrelar incertezas ao modelo e às inferências. A regressão linear bayesiana demonstrou resultados compatíveis com os obtidos através da abordagem clássica para os parâmetros de interesse.

Para uma análise futura, pode-se considerar realizar as seguintes análises: Analisar o impacto ao alterar o número de cadeias de markov e de amostras retiradas da distribuição posterior; Analisar diferentes métodos de Monte Carlo via cadeia de Markov; Analisar a influência da quantidade de dados observados na inferência bayesiana. Por fim, propõe-se como trabalho futuro a realização de uma regressão linear múltipla.

REFERÊNCIAS

BECKER, S.; PLUMBLEY, M. Unsupervised neural network learning procedures for feature extraction and classification. Springer, 1996.

DOMINGOS, P. A few useful things to know about machine learning. 2012. Disponível em: <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>.

FEIGELSON, E. D.; BABU, G. J. **Modern Statistical Methods for Astronomy**. [S.l.: s.n.], 2012. ISBN 9780521767279.

FOUNDATION, A. P. S. **About Sloan Digital Sky Survey**. 2018. Disponível em: <https://sloan.org/programs/science/sloan-digital-sky-survey>. Acesso em: jun. 2018.

FUKUGITA, M.; ICHIKAWA, T.; GUNN, J. E.; DOI, M.; SHIMASAKU, K.; SCHNEIDER, D. P. **The Sloan Digital Sky Survey Photometric System**. American Astronomical Society, 1996. Disponível em: <http://adsbit.harvard.edu//full/1996AJ....111.1748F/0001748.000.html>.

J., S.; T.V., W.; C., F. **Probabilistic programming in Python using PyMC3**. [S.l.]: PeerJ Computer Science 2:e55, 2016.

KAELBLING, L. P.; LITTMAN, M. L.; MOORE, A. W. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 1996.

KEPLER, S. O.; SARAIVA, M. F. O. **Astronomia e Astrofísica**. Departamento de Astronomia - Instituto de Física - UFRGS, 2014. Disponível em: <http://astro.if.ufrgs.br/livro.pdf>. Acesso em: jun. 2018.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. ISBN 0070428077.

NEWMAN, M. E. J.; BARKEMA, G. T. **Monte Carlo Methods in Statistical Physics**. [S.l.]: Oxford University Press, 1999.

NEYMAN, J. Outline of a theory of statistical estimation based on the classical theory of probability. Royal Society of London, 1937.

NILSSON, N. J. **The Quest for Artificial Intelligence**. [S.l.: s.n.], 2009. ISBN 9780521122931.

REIS, M. M. **INFERÊNCIA ESTATÍSTICA – Estimação de Parâmetros**. 2018. Disponível em: <https://www.inf.ufsc.br/~marcelo.menezes.reis/Cap9.pdf>. Acesso em: jun. 2018.

STERKEN, C.; MANFROID, J. **Astronomical photometry: a guide, Astrophysics and space science library**. [S.l.]: Springer, 1992.

WASSERMAN, L. **All of statistics: a concise course in statistical inference**. [S.l.]: Springer, 2004.