MSc

2.º CICLO

FCUP
2019/2020

U.PORTO

Regressão Linear

Octávio Pedro Fernandes da Silva

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Regressão Linear

Octávio Pedro Fernandes da Silva

Dissertação de Mestrado apresentada à Faculdade de Ciências da Universidade do Porto em
Ensino de Matemática no 3º Ciclo do Ensino Básico e no Secundário
2019/2020

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Regressão Linear

Mestrado em Ensino de Matemática no 3º Ciclo do Ensino Básico e no Secundário

Orientadores científicos:

Professora Maria João Costa, Faculdade de Ciências da Universidade do Porto

Professora Maria João Rodrigues, Faculdade de Ciências da Universidade do Porto

Cooperante:

Professora Dilma Tuna, Escola Secundária João Gonçalves Zarco, Matosinhos

Resumo: Regressão Linear. Neste relatório, pretende-se construir e trabalhar os modelos de regressão linear simples e múltipla.

O relatório inicia-se com a minha reflexão global acerca do Mestrado em Ensino da Matemática no 3º Ciclo e no Ensino Secundário, e também, de como decorreu o meu estágio na Escola João Gonçalves Zarco no ano letivo 2019/2020.

Este trabalho divide-se em duas partes estruturantes. Na primeira é elaborada uma parte didática para alunos do 11º ano, com o objetivo de estes aplicarem o método dos mínimos quadrados com 2 e 3 variáveis num problema de contexto real. Nesta parte, são ainda introduzidas as funções reais de variável vetorial, para que os estudantes entendam melhor como funciona o método dos mínimos quadrados aplicado a 3 variáveis.

Na segunda parte é dado um exemplo com 2 variáveis, que serve como motivação para a construção e análise do método dos mínimos quadrados. Enquadrado no mesmo exemplo, mas agora com 3 variáveis, introduz-se o método dos mínimos quadrados multivariado com a utilização de matrizes. Por último, mostra-se porque o método dos mínimos minimiza os desvios verticais das estimativas dos modelos lineares estudados.

Palavras-chave: variável**,** método dos mínimos quadrados, regressão linear simples, regressão linear múltipla, aplicações.

Abstract: Linear Regression. In this report, it is intended to build and work the simple and multiple linear regression models.

The report begins with my global reflection on the Master's degree in Mathematics Teaching in the 3rd Cycle and Secondary Education, and also, how my internship at João Gonçalves Zarco School took place in the academic year 2019/2020.

This work is divided into two structuring parts. In the first, a didactic part is elaborated for 11th grade students, with the objective of applying the least squares method with 2 and 3 variables in a real context problem. In this part, the actual functions of vector variable are also introduced, so that students understand better how the least squares method applied to 3 variables works.

In the second part is given an example with 2 variables, which serves as motivation for the construction and analysis of the least squares method. Framed in the same example, but now with 3 variables, the multivariate least squares method is introduced with the use of matrices. Finally, it is shown why the least squares method minimizes the vertical deviations of the estimates of the linear models studied.

Keywords: variable, least squares method, simple linear regression, multiple linear regression, applications.

Conteúdo

INTRODUÇÃO

Este trabalho mais do que o fim de mais uma etapa da minha vida foi o início de uma etapa de aprendizagem constante e infinita no mundo do ensino.

O tema que decidi desenvolver no meu relatório de estágio foi a regressão linear, por considerar este um assunto muito interessante a que na minha opinião não é dado o devido valor. A razão pela qual eu considero que o tema carece de algum foco, é porque por norma este é lecionado no último capítulo, a estatística, do programa nacional de estudos do 11º ano. Como o programa é tão extenso, acontece que os professores ficam com pouco tempo para falar devidamente acerca da regressão linear, trabalhando este assunto muito aquém das suas potencialidades.

Este relatório está estruturado em 3 partes. Na primeira refleti acerca de como decorreu o ano letivo de 2019/2020 em que estive colocado como professor estagiário na escola secundária João Gonçalves Zarco em Matosinhos. Nesta parte abordei ainda os aspetos positivos e negativos do estágio, do que gostei e não gostei, a minha evolução, o que aprendi e o que tenho de melhorar, bem como alguns aspetos do ensino que a meu ver deveriam ser reformulados.

Na segunda parte do relatório desenvolvi uma atividade didática relacionada com o tema deste relatório, que se destina a alunos de Matemática que tenham concluído ou estejam prestes a concluir o 11º ano. A atividade didática divide-se em três pontos. No primeiro aplica-se o método dos mínimos quadrados com 2 variáveis num problema real em que os alunos têm de efetuar medições e chegar a conclusões. No outro ponto da atividade didática introduzi muito superficialmente funções reais de variável vetorial, para que faça sentido finalizar a atividade didática com uma aplicação do método dos mínimos quadrados para 3 variáveis numa situação real.

A terceira e última parte do relatório foi dedicada à parte científica onde explorei detalhadamente a regressão linear, o contexto de aplicação da mesma, a demonstração do modelo, a posterior análise do enquadramento do modelo linear, bem como alguns outros resultados e observações.

REFLEXÃO

O ano escolar 2019/2020 teve início no dia 18 de setembro na escola João Gonçalves Zarco, onde tive a oportunidade de estagiar com a orientação da professora cooperante Dilma Tuna. A organização dos períodos letivos nesta escola é bastante diferente do que acontece na maioria das outras, o ano escolar divide-se em dois semestres, em que é feita uma avaliação dos estudantes no final de cada semestre.

Cada semestre divide-se em dois períodos existindo uma pausa de três dias entre eles, nesta interrupção não existe avaliação dos estudantes. Esta diferente organização do ano escolar, do meu ponto de vista, tem as suas vantagens bem como as suas desvantagens, considero que para os professores é muito melhor pois têm de dar notas apenas 2 vezes por ano, em janeiro e no fim do ano letivo. Como nesta organização a primeira avaliação é realizada mais tarde, será assim a meu ver mais sensata e refletida. Outro fator que também contribui para uma avaliação mais ponderada, é o facto de as reuniões de avaliação serem realizadas durante as interrupções letivas, pois os professores estão menos ocupados podendo assim ponderar melhor a avaliação de cada estudante.

Por outro lado, estas interrupções letivas têm uma influência negativa no estudo de alguns alunos, pois estes vêm estas pequenas pausas como minifférias onde não têm de estudar, o que põe em causa o seu progresso no estudo. Estou de acordo com esta estruturação do ano letivo, mas penso que as pausas letivas deveriam ser ponderadas, na medida do seu benefício.

As turmas atribuídas à professora Dilma Tuna foram duas do sétimo ano, o 7º 1 e o 7º 2, e duas do décimo primeiro ano, o 11º 1 e o 11º7. A experiência que adquiri com estas turmas foi muito diversificada. Por serem turmas todas muito distintas umas das outras, contribuíram para alargar o meu campo de estratégias e de conhecimentos no mundo do ensino.

Com as turmas do 7º ano percebi que ensinar vai muito para além de conhecimentos matemáticos. Principalmente na turma do 7º 1 existiam muitos alunos com dificuldades de aprendizagem e de concentração, bem como um comportamento muito pouco adequado para a sala de aula. A primeira aula que lecionei a esta turma fez-me perceber que a atenção de um professor na sala de aula tem de ser a 100%, este deve estar atento a tudo e a todos. Numa sala de aula acontecem bastantes situações distintas, que o professor tem de ter em conta e agir em conformidade. Esta turma fez-me aprender a lidar com alunos que estão sempre agitados e com pouca

vontade de trabalhar, conseguindo assim controlar e manter a ordem numa sala de aula com alunos com estas características.

Considero que foi fundamental para a minha progressão no ensino contactar com turmas do 7º ano, pois sem esta experiência e a orientação da professora Dilma para conseguir lidar com estes estudantes, no futuro iria ter dificuldade a ensinar a turmas do básico. Na minha opinião, se possível, todos os novos estagiários de matemática deveriam ser colocados numa turma do 7º ano.

Em relação às turmas do 11º ano, estas não poderiam ser mais distintas. O 11º 1 era maioritariamente constituído por alunos de excelência, em que para entrar nesta turma era necessário ter notas acima da média e ser entrevistado, de modo a selecionar apenas alunos com o objetivo de ingressar em cursos de prestígio no ensino superior. Esta foi a turma que me deu mais vontade e motivação para lecionar, visto que os estudantes eram bastante trabalhadores e interessados. Nesta turma existia um ótimo ambiente de trabalho, o que me permitia ao apresentar novos conteúdos fazer as respetivas demonstrações. Os estudantes gostavam das demonstrações, pois assim percebiam de onde surgiam os conteúdos, e que as propriedades e os teoremas da matemática tinham uma origem e não eram simplesmente "caídos do céu".

Pelo contrário, na turma do 11º7 os estudantes eram desinteressados e desmotivados, poucas eram as vezes em que os alunos realizavam as tarefas propostas para casa, o que comprometia bastante o trabalho diário e a sua aprendizagem. Nesta turma existia um caso particular de um aluno que a tinha negativa do 10º ano, e acrescentava o problema de não querer ser ajudado. Com a indicação da professora Dilma, eu e o meu colega de estágio Nuno Ferreira nos primeiros meses tentamos motivar este aluno, sentávamo-nos alternadamente na mesma secretária que o aluno com o objetivo de explicar melhor a matéria e ajudá-lo a resolver os exercícios. Este acompanhamento em nada resultou, pois, o aluno não colaborava, tinha de ser forçado a trabalhar, o que não trazia resultados. Este caso fez-me perceber que num nível de ensino do 11º ano o professor algumas vezes tem de ponderar quanta atenção deve dar a um aluno, porque a turma pode acabar por ser prejudicada.

A diversidade de estudantes com que contactei e as grandes diferenças de atitudes das turmas foram uma mais valia para mim como futuro professor. Esta diversidade exigiu que eu optasse por diferentes tipos de estratégia de ação, para arranjar forma de conseguir manter a ordem na sala de aula e conseguir proporcionar um bom ambiente, propício à aprendizagem.

Ao longo destes meses a acompanhar a professora Dilma diariamente, senti a verdadeira importância do professor. Para fazer um bom trabalho e de facto ensinar Matemática, o professor tem de fazer muito mais do que simplesmente lecionar a matéria, tem de ser um elo entre o conhecimento e os estudantes. O que mais evidenciei na postura da professora Dilma, foi a sua diversificação e adaptabilidade a cada situação, agindo com diferentes estratégias de turma para turma, matéria para matéria e aluno para aluno. Senti que a minha evolução em termos de adaptação a cada situação foi gradualmente melhorando, percebendo ao longo do ano qual a melhor forma de agir em determinado momento.

Relativamente ao programa nacional de estudos, no 7º ano na minha opinião, enquadra-se perfeitamente, sendo simples e claro. Considero apenas que o tema de figuras geométricas está muito extenso e confuso, defendo assim que este tema deveria ser reestruturado.

No 11º ano penso que o plano de estudo é demasiado extenso, comprovado pelo facto de que muitas vezes este não ser todo lecionado, o que acaba por prejudicar as aprendizagens do 12º ano, pois o professor tem de ensinar o que não foi ensinado pela falta de tempo. Tudo isto complica ainda mais a tarefa do professor do 12º ano, visto que o programa deste ano também ele é muito extenso e agrava mais por ser um ano de exame nacional.

A escola secundária João Gonçalves Zarco é uma escola que inclui alunos muito distintos, uma vez que esta abrange uma grande área metropolitana, o município de Matosinhos. Considero a escola excelente em todos os parâmetros, a nível de transportes, existe uma grande afluência de metros e autocarros com paragens muito próximas da escola. As instalações da escola são muito recentes o que proporciona um ambiente muito agradável, tornando o exercício de apreender e ensinar muito mais reconfortante. Os profissionais que lá trabalham são todos muito prestáveis, desde os professores aos funcionários, que me acolheram a mim e ao meu colega Nuno da melhor maneira. É de salientar a biblioteca da escola, com ótimas instalações, um arquivo vasto, diversificado e atualizado. Apesar destas excelentes condições, constatei que durante o ano letivo vi poucos estudantes a usufruir do que este espaço tem para oferecer.

Foi gratificante ser colocado como estagiário da professora Dilma Tuna, visto ser uma professora com uma sólida e vasta experiência de ensino, conseguindo assim transmitir os melhores valores do ensino a mim e ao meu colega Nuno. A professora tem bastante experiência na formação dos seus estagiários, visto já o ter feito várias

vezes. A professora criou um ótimo ambiente para o meu estágio, dando-me a liberdade para lecionar as aulas que quisesse e me sentisse mais confortável, o que me motivou muito, pois pude lecionar os conteúdos que mais gosto. A professora deu-me autonomia total para lecionar as aulas, corrigindo e melhorando os planos de aula feitos por mim, o que contribuí imenso para a minha confiança durante as minhas prestações.

As professoras orientadoras foram muito prestáveis tanto no estágio como no desenvolvimento do relatório, constantemente me motivavam para eu fazer um bom trabalho e avançar desde cedo na construção do relatório não lhe dando menor importância. Ter duas professoras orientadores foi ótimo pois permitiu-me ter opiniões diferentes e diversificadas, visto que cada uma era de uma área científica diferente. Gostei muito de ter as professoras Maria João Costa e Maria João Rodrigues como orientadoras e queria agradecer pela forma como me trataram e por todo o apoio prestado.

Em relação ao Mestrado em Ensino da Matemática para Professores do 3º ciclo e Ensino Secundário, considero que o primeiro ano deveria ser reestruturado pois senti-me desmotivado. Na minha opinião o trabalho que realizei nesse ano não contribui muito para a minha formação enquanto professor de Matemática. Já o segundo ano foi excelente para mim, um ano exigente, mas em que pude vivenciar diariamente o que era o ensino em Portuga. Pude pôr em prática os conhecimentos e técnicas da metodologia de ensino que conhecia, consolidá-los e aprender outras técnicas durante prática da docência.

O covid-19 teve implicações muito fortes e rápidas na vida de toda a gente do mundo, uma experiência surreal para todos, que ninguém imaginaria viver. Foi muito marcante ver de perto o que um pequeno vírus pode fazer a uma sociedade e perceber a sua fragilidade. Quando se deram os primeiros casos em Portugal é que as pessoas começaram a levar a sério este pequeno microrganismo, na escola o pânico era geral, todos os que a frequentavam tinham uma sensação de medo e suspeita uns dos outros, pois estávamos a enfrentar algo que não sabíamos de onde poderia vir. Os professores seguiam à risca as indicações que eram dadas, tentando passar a informação da melhor maneira para os estudantes e de forma a não causar mais pânico. Decretado o estado de emergência e a obrigação de confinamento, os professores tiveram de se consciencializar de que os alunos não voltariam mais a escola este ano. Para que o ensino não estagnasse os professores foram obrigados a arranjar novas ferramentas para poderem trabalhar à distância com os estudantes. Foi uma experiência muito desafiante o ensino à distância. Adaptar os conteúdos a aulas online é muito exigente

sendo necessário simplificar e clarificar ao máximo, visto que à distância há mais dificuldades em passar as ideias de forma esclarecedora. Sendo assim, o programa não pôde ser cumprido, o nível de exigência baixou imenso e as avaliações foram muito mais simples. Neste período de quarentena todos tivemos tempo para refletir acerca de tudo um pouco, penso que todos chegaram à conclusão da vulnerabilidade da vida, de que existem diversas ameaças no mundo que não estamos prontos para lidar. Com estes acontecimentos concluí que é necessário termos um espírito bastante aberto e uma autodisciplina exigente, para quando as dificuldades surgirem cada um de nós tenha motivação e capacidade para as superar.

Este foi o ano mais importante para mim como futuro professor de matemática, foi o ano em que comecei a construir a minha identidade docente, no fundo autoconstruir o professor que quero ser. Percebi que esta autoconstrução não é fácil, pois a imagem que cada um tem de si próprio muitas vezes não corresponde à realidade, os conteúdos que queremos ensinar e o modo como o queremos fazer não acontece como desejamos. Os objetivos que estão definidos no papel e na minha mente para lecionar uma aula, na prática não são muitas vezes atingidos como realmente gostaria. No fundo percebi que se quero ser um professor de excelência, um professor a 100%, tenho de fazer um trabalho muito superior, um trabalho a 1000 %.

CAPÍTULO I- ATIVIDADE DIDÁTICA: MÉTODO DOS MÍNIMOS QUADRADOS

1.1. INTRODUÇÃO, PÚBLICO ALVO, OBJETIVOS

A atividade didática está dividida em três. A primeira parte consiste em estudar um exemplo prático da aplicação do método dos mínimos quadrados, num problema onde existe a necessidade de estimar um parâmetro. A segunda parte da atividade tem dois objetivos, incentivar os alunos a alargar os seus horizontes da matemática e prepará-los para o último exercício didático. Assim, introduzi de forma muito simples funções reais de variável vetorial, para melhor entenderem o novo conceito da aplicação do método dos mínimos quadrados quando existem 3 variáveis, uma variável que depende das outras duas, ou seja, uma função $f: \mathbb{R}^2 \to \mathbb{R}$. Relativamente às funções reais de variável vetorial os alunos irão contactar com alguns exemplos e propriedades muito simples.

Por último é introduzido o algoritmo do método dos mínimos quadrados quando temos uma variável que depende de duas, estabelecendo-se aqui a ligação com as funções $f: \mathbb{R}^2 \to \mathbb{R}$. Os estudantes terão de aplicar o método dos mínimos quadrados com 3 variáveis num exemplo prático, onde os dados são fornecidos.

Na minha opinião considero pertinente introduzir outro tipo de funções, mesmo que seja de forma muito simples com alguns exemplos. Isto pelo facto de os alunos no secundário estarem habituados, mas sem perceber o porquê, a ouvir os professores nomear uma função dizendo “dado f uma função real de variável real, que podem abreviar por f.r.v.r”. O que me deixa alarmado é os alunos não se questionarem acerca da natureza da função, acabando por aceitar o que a professora diz sem se questionarem: Será que existe outro tipo de funções?

A atividade tem como público alvo os estudantes que já terminaram a disciplina de matemática do 11º ano ou que estejam prestes a terminá-la, pois esta atividade envolve principalmente o método dos mínimos quadrados, tema este que normalmente é abordado no último capítulo do programa, a estatística. Certamente, os alunos do curso de ciências irão estar mais motivados para a resolução da primeira parte que envolve medições de massas e volumes, e o conceito de densidade, visto que estes estudantes têm a disciplina de física onde normalmente trabalham estes conceitos. Os estudantes de outras áreas não terão qualquer desvantagem na resolução deste exercício, pois os conceitos estão explicitados.

Esta atividade seria para ser aplicada num contexto de sala de aula, acompanhada por um professor de Matemática, para que seja possível tirar dúvidas e esclarecer qualquer questão que surja a algum dos estudantes. A atividade tem também como objetivo que os alunos explorem um pouco as funcionalidades da calculadora. Em

particular, é necessário efetuar o cálculo de vários somatórios, que com a utilização de listas na calculadora se tornam muito mais simples. No livro do 11º “Novo Espaço” da Porto Editora, existe o tutorial de como fazer cálculo de somatórios com listas que os alunos devem consultar. O professor também deverá ajudar os alunos a perceber como trabalhar estes conceitos com a calculadora gráfica.

1.2. APLICAÇÃO PRÁTICA DO MÉTODO DOS MÍNIMOS QUADRADOS: PROBLEMA DO SENHOR JOAQUIM

O senhor Joaquim é proprietário de uma frutaria, e nos últimos tempos os seus clientes queixam-se das laranjas, afirmando que não são boas, pois não têm muito sumo.

O senhor Joaquim decidiu assim mudar de fornecedor de laranjas, pois sabia que normalmente as laranjas que possuíam mais sumo eram as mais densas. Tinha diferentes fornecedores de laranjas, mas como estava decidido a agradar os clientes, queria escolher o fornecedor com as laranjas mais densas. Para isso teria de realizar um estudo acerca da densidade das laranjas de cada um dos diferentes fornecedores. Como o estudo não ia ser assim tão fácil, decidiu pedir ajuda ao seu filho João, que frequentava o 11º ano do curso de ciências, e que por isso tinha bons conhecimentos de matemática e físico-química.

O João iria desenvolver um estudo com 8 laranjas escolhidas aleatoriamente de cada um dos fornecedores, para perceber qual dos fornecedores tinhas as laranjas mais densas. Para o estudo o João precisava da massa e o volume de cada laranja, pois sabia que a densidade pode ser calculada pelo quociente entre a massa e o volume.

Para medir a massa das laranjas em gramas o João utilizou uma balança de cozinha.

Para o volume o João iria mergulhar cada uma das laranjas num recipiente com água, com marcações volumétricas, e através da deslocação da água medir o volume das laranjas. Nesta etapa o João teve dificuldades, visto que as laranjas quando mergulhadas em água flutuavam, o que impedia o cálculo do seu volume. Refletindo melhor, a solução para o problema seria obter o volume de cada laranja por uma aproximação ao volume da esfera, para isso o João mediu o diâmetro de cada laranja.

Etapas do estudo do João para cada fornecedor:

1. Pesou e mediu o diâmetro de cada uma das 8 laranjas, em gramas e centímetros;
2. Calculou uma aproximação do volume de cada laranja a uma esfera em centímetros cúbicos, com arredondamento às unidades;
3. Calculou a densidade de cada laranja, conservando 3 casas decimais;
4. Calculou a densidade média das 8 laranjas;
5. Representou os pontos volume/massa num referencial;

6. Aplicou, manualmente, o método dos mínimos quadrados para estimar a reta que passa nos pontos do referencial representado na etapa anterior;
7. Adicionou o ponto $(0,0)$ aos pontos representados na etapa 5;
8. Aplicou novamente de forma manual o método dos mínimos quadrados, para estimar a reta que passa no novo conjunto de pontos;

1.2.1. ESTUDO DO JOÃO

Deves agora simular o estudo que o João realizou para um dado fornecedor.

Para isso deves:

- Escolher aleatoriamente 8 laranjas provenientes do mesmo fornecedor;
- Realizar as etapas do estudo do João;
- Completar a tabela abaixo;
- Utilizar o método dos mínimos quadrados descrito e apresentar cada um dos passos que realizares para a obtenção das duas retas dos mínimos quadrados;
- Representar numa folha quadriculada, com cores distintas, as retas e o conjunto de pontos obtidos num referencial cartesiano adequado;
- Responder, justificando, a cada uma das seguintes questões.

Questões:

1. Qual a fórmula que relaciona o diâmetro de uma esfera e o seu volume?

2. As densidades das 8 laranjas escolhidas são muito distintas? Será uma solução plausível para o problema do senhor Joaquim comparar a densidade média das 8 laranjas escolhidas dos diferentes fornecedores?

3. O declive da reta dos mínimos quadrados obtida na etapa 6 do estudo do João não será também um ponto de partida razoável para a comparação da densidade das laranjas dos diferentes fornecedores?

4. No contexto do problema faz sentido o que o João fez na etapa 7, adicionar o ponto $(0,0)$? Dá uma explicação do que significa adicionar o ponto $(0,0)$ e averigua o que acontece a reta dos mínimos quadrados com a adição desse ponto.

5. Se aplicarmos o método dos mínimos quadrados na calculadora obténs os mesmos resultados que obtiveste à mão? Para isso deves aplicar o método dos mínimos quadrados na calculadora aos pontos das etapas 5 e 7.

Massa (gramas)								
Diâmetro (centímetros)								
Volume (centímetros cúbicos)								
Densidade (gramas por centímetros cúbicos)								

Tabela 1- Tabela a utilizar pelos estudantes ao aplicarem o Estudo do João.

DESCRIÇÃO DO MÉTODO DOS MÍNIMOS QUADRADOS

Dado um conjunto de n pontos, $n \epsilon \mathbb{N}$, do plano $\{(x_1, y_1), \dots, (x_n, y_n)\}$, a reta obtida aplicando o método dos mínimos quadrados será a reta de equação $y = ax + b$ que minimiza as somas dos quadrados das distâncias de cada um dos pontos à reta.

Seja $x = \{x_1, \dots, x_n\}$.

Seja $y = \{y_1, \dots, y_n\}$.

O algoritmo é dado por:

1. Calcular a média de x;
2. Calcular a média de y;
3. Calcular $\sum_{i=1}^{n} {x_i}^2$;
4. Calcular $SS_x = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$;
5. Calcular $\sum_{i=1}^{n} x_i\, y_i$;

A reta é dada por:

$$a = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{SS_x}$$

$$b = \bar{y} - a\bar{x}$$

1.2.2. SIMULAÇÃO DO ESTUDO DO JOÃO: RESOLUÇÃO

Massa (gramas)	200	203	207	227	257	265	269	275
Diâmetro (centímetros)	7,8	7,9	8	8,2	8,4	8,5	8,6	8,8
Volume (centímetros cúbicos)	248	258	268	289	310	322	333	357
Densidade (gramas por centímetros cúbicos)	0,806	0,787	0,772	0,785	0,829	0,823	0,808	0,770

Tabela 2- Tabela a utilizar pelos estudantes ao aplicarem o Estudo do João (Resolução).

$$Densidade\ média \rightarrow \rho_m$$

$$\rho_m = \frac{0{,}806 + 0{,}787 + 0{,}772 + 0{,}785 + 0{,}829 + 0{,}823 + 0{,}808 + 0{,}770}{8} = 0{,}798\ g/cm^3$$

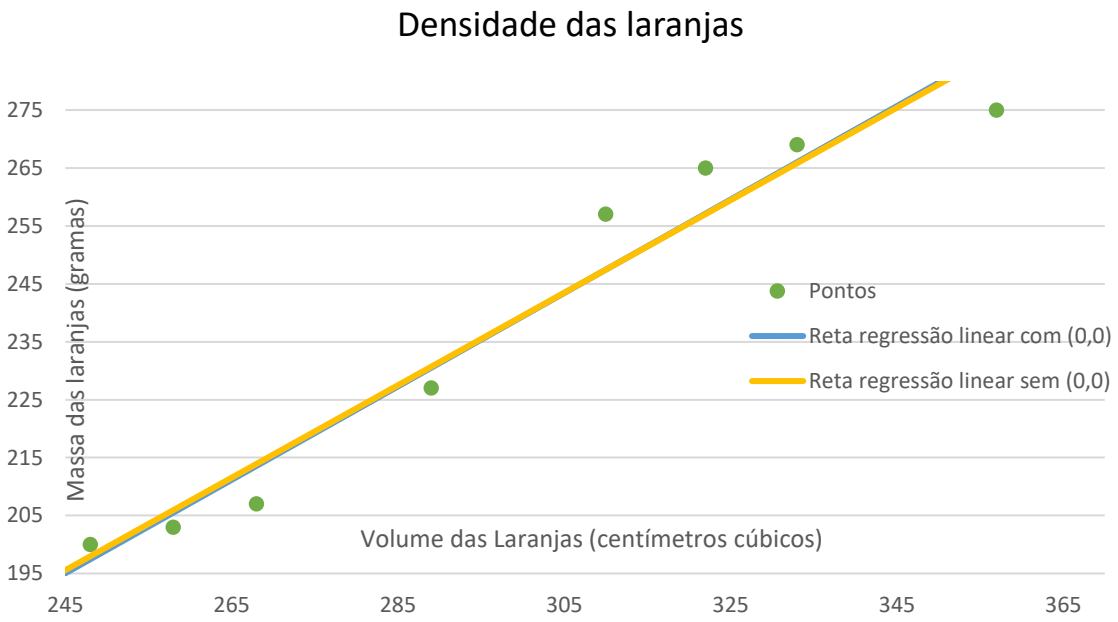
Reta dos mínimos quadrados obtida utilizando os pontos da tabela:

$$y = 0{,}806x - 2{,}544$$

Reta dos mínimos quadrados obtida utilizando os pontos da tabela com $(0{,}0)$:

$$y = 0{,}799x - 0{,}265$$

Representação do conjunto de pontos (volume/massa das laranjas) e das duas retas obtidas com a aplicação do modelo de regressão linear, com e sem o ponto (0,0).



Resposta as questões:

1. Qual a fórmula que relaciona o diâmetro de uma esfera e o seu volume?

$$d \rightarrow diâmetro$$

$$r \rightarrow raio$$

$$d = 2r \iff r = \frac{d}{2}$$

$$V_{esfera} = \frac{4}{3}\pi r^3 = \frac{4}{3}\pi\left(\frac{d}{2}\right)^3 = \frac{4}{3}\pi\frac{d^3}{8} = \frac{\pi}{6}d^3$$

2. As densidades das 8 laranjas escolhidas são muito distintas? Será uma solução plausível para o problema do senhor Joaquim comparar a densidade média das 8 laranjas escolhidas dos diferentes fornecedores?

$$variação\ da\ densidade = \rho_{máxima} - \rho_{mínima} = 0{,}829 - 0{,}770 = 0{,}059$$

As densidades das 8 laranjas não são muito diferentes, a laranja mais densa tem uma densidade de $0{,}829$ e a menos densa de $0{,}770$ uma diferença de apenas $0{,}059$. Acho que será uma solução possível para o problema do senhor Joaquim considerar a densidade média das 8 laranjas, e escolher o fornecedor com a maior densidade média das laranjas estudadas.

3. O declive da reta dos mínimos quadrados obtida na etapa 6 do estudo do João não será também um ponto de partida razoável para a comparação da densidade das laranjas dos diferentes fornecedores?

O declive calculado na etapa 6 foi de $0{,}806$ e o valor da densidade média das 8 laranjas é de $0{,}798$, a diferença entre estes valores é de apenas $0{,}008$. Este valor é bastante pequeno, portanto se consideramos que a densidade média é uma possível solução então o declive da reta dos mínimos quadrados obtida na etapa 6 do estudo do João também será uma solução admissível para o problema do senhor Joaquim.

4. No contexto do problema dá a tua opinião acerca se faz ou não sentido fazer o que o João fez na etapa 7, adicionar o ponto $(0{,}0)$? Dá uma explicação do que significa adicionar o ponto, e averigua o que acontece a reta dos mínimos quadrados com a sua adição aos valores da tabela.

Faz todo o sentido no contexto do problema adicionar o ponto $(0{,}0)$ pois uma laranja que tenha volume 0 terá obviamente massa 0.

A nova reta dos mínimos quadrados irá enquadrar-se melhor no problema, pois a ordenada na origem fica mais próxima da origem do referencial, passa de $-2{,}544$ para $-0{,}265$. Isto favorece a situação de uma laranja ter volume nulo, visto que a sua massa também deveria ser nula.

O declive da reta dos mínimos quadrados também melhora no contexto do problema pois como o declive pode ser visto como uma aproximação da densidade das laranjas daquele fornecedor em estudo. Com a adição do ponto $(0{,}0)$ o declive da reta fica mais próximo da densidade média que é de $0{,}798$, passando de um declive de $0{,}806$ para $0{,}799$, sendo agora a diferença entre o declive e a densidade média de apenas $0{,}001$.

5. Se aplicarmos o método dos mínimos quadrados na calculadora obténs os mesmos resultados que obtiveste à mão? Para isso deves aplicar o método dos mínimos quadrados na calculadora aos pontos das etapas 5 e 7.

Sim obtemos exatamente os mesmo resultados com 3 casas decimais.

1.3. PARA LÁ DA FUNÇÃO REAL DE VARIÁVEL REAL

Desde o 7º ano tens vindo a falar e a estudar funções, um dos principais e mais importantes temas da Matemática. É de notar que dentro do universo da Matemática as funções ocupam um significante espaço, uma galáxia, e dentro desta galáxia tu apenas tens vindo a estudar um sistema solar. Nesta atividade, terás a oportunidade de conhecer um novo sistema solar e assim alargar os teus horizontes.

As funções que conheces são denominadas de funções reais de variável real, "f.r.v.r", que no fundo representam uma correspondência entre números. Por exemplo a função $f(x) = x^2$, é a função que ao número x faz corresponder o seu quadrado. Para podermos caracterizar corretamente a função temos de apresentar o seu domínio e o conjunto de chegada. A função f ficará corretamente definida da forma:

$$f: \mathbb{R} \to \mathbb{R}$$

$$x \longmapsto x^2$$

Podemos pensar no conjunto $\mathbb{R}$, dos números reais como uma reta, que apenas tem 1 dimensão, assim o conjunto dos números reais terá também 1 dimensão.

Existem espaços com mais dimensões, como é exemplo o plano, que tem 2 dimensões. Definimos o plano como $\mathbb{R}^2$. Já o espaço tem 3 dimensões, que será assim definido como $\mathbb{R}^3$.

As funções que conheces, estabelecem correspondências entre espaços de uma dimensão, mas existem outro tipo de funções que estabelecem correspondências entre espaços de diferentes dimensões e com mais de uma dimensão. As funções que iremos analisar a seguir são funções que estabelecem correspondências entre espaços de duas dimensões com espaços de apenas uma dimensão. Alguns exemplos:

$$g: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto x + y$$

$$h: \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$$

$$(x, y) \longmapsto x^y$$

$$f: D_f \to \mathbb{R}$$

$$(x, y) \longmapsto \frac{x}{y}$$

Estes tipos de funções denominam-se por funções reais de variável vetorial, pois a um vetor do plano, a função faz corresponder um número.

Estas funções tal como as funções reais de variável real, têm um domínio e um contradomínio, podem ser injetivas e/ou sobrejetivas.

1.3.1. PARA LÁ DA FUNÇÃO REAL DE VARIÁVEL REAL: EXERCÍCIOS

Sejam f, g e h as funções indicadas acima.

1. Completa os espaços em branco:

1.1. $g\,(1,2) =$

1.2. $g\,(\ ,\ ) = 4$

1.3. $h\,(1,2) =$

1.4. $h\,(\ ,\ ) = 8$

1.5. $f\,(1,1) =$

1.6. $f\,(2,2) =$

2. No exercício anterior calculaste $f(1,1)$ e $f(2,2)$ e reparaste que eram iguais. Podemos concluir que a função f não é injetiva, pois existem objetos diferentes $(1,1) \neq (2,2)$ que têm a mesma imagem $f(1,1) = f(2,2)$.

   2.1. Com um raciocínio semelhante prova que as funções g e h também não são injetivas.

3. Dá um exemplo de uma função $j: \mathbb{R}^2 \to \mathbb{R}$.

4. Existe algum ponto de $\mathbb{R}^2$ que não tem imagem? Será que o domínio da função $f\ (D_f)$ pode ser $\mathbb{R}^2$?

1.3.2. PARA LÁ DA FUNÇÃO REAL DE VARIÁVEL REAL: EXERCÍCIOS-RESOLUÇÃO

Sejam g, h e f as funções indicadas acima.

$$g: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto x + y$$

$$h: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto x^y$$

$$f: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto \frac{x}{y}$$

1. Completa os espaços em branco:

   1.1. $g(1,2) = 1 + 2 = 3$

   1.2. $g(2,2) = 4$

   1.3. $h\,(1,2) = 1^2 = 1$

   1.4. $h\,(2,3) = 8$ por exemplo, existem outras soluções.

   1.5. $f(1,1) = \frac{1}{1} = 1$

   1.6. $f(2,2) = \frac{2}{2} = 1$

2. No exercício anterior calculaste $f(1,1)$ e $f(2,2)$ e reparaste que eram iguais, podemos concluir que a função f não é injetiva, pois existem objetos diferentes $(1,1) \neq (2,2)$ que correspondem a mesma imagem $f(1,1) = f(2,2)$.

2.1. Com um raciocínio semelhante prova que as funções g e h também não são injetivas.

g não é injetiva, pois existem objetos diferentes $(1,2) \neq (2,1)$ que correspondem à mesma imagem $g(1,2) = g(2,1) = 3$

h não é injetiva, pois existem objetos diferentes $(1,2) \neq (1,1)$ que correspondem à mesma imagem $h(1,2) = h(1,1) = 1$

3. Dá um exemplo de uma função $j: \mathbb{R}^2 \to \mathbb{R}$.

$$j: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto x$$

4. Existe algum ponto de $\mathbb{R}^2$ que não tem imagem? Será que o domínio da função f (D_f) pode ser $\mathbb{R}^2$?

Os pontos da forma $(x, 0)$, com $x \in \mathbb{R}$, não tem imagem pela função f pois não faz sentido dividir um número por 0. Assim, como não podemos calcular $f(x, 0)$ para nenhum valor de $x \in \mathbb{R}$ temos que o maior subconjunto onde a função f pode estar definida é $\mathbb{R}^2 \setminus \{(x,0)\}$.

1.4. MÉTODO DOS MÍNIMOS QUADRADOS APLICADO A 3 VARIÁVEIS

No capítulo de estatística do 11º ano tens trabalhado o conceito da reta dos mínimos quadrados, uma reta que relaciona duas variáveis representadas por x e y através de um modelo linear. Para construir o modelo linear temos como base n observações de cada uma das variáveis $\{(x_1, y_1), \dots, (x_n, y_n)\}$. Podemos ver cada par de observações como um ponto do plano.

O nosso objetivo é encontrar uma relação linear entre as variáveis x e y, da forma $y = ax + b$. A variável y é designada como dependente e a variável x como independente. Se pensarmos em termos de funções, a reta dada pelos mínimos quadrados será uma função real de variável real da forma:

$$y: \mathbb{R} \to \mathbb{R}$$

$$x \longmapsto ax + b$$

Pensemos agora no caso de ter de três variáveis x, y e z e n observações destas, $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$. Podemos ver cada uma das observações como um ponto do espaço.

Será que também se pode obter um modelo linear que relacione estas variáveis?

Sim, se designarmos a variável z como dependente e as variáveis x e y como independentes, podemos ver então a variável z como uma função que depende das variáveis x e y. Temos aqui uma função real de variável vetorial, pois a cada par (x, y) corresponde um valor da variável z.

$$z: \mathbb{R}^2 \to \mathbb{R}$$

$$(x, y) \longmapsto z(x, y)$$

Ao aplicarmos o método dos mínimos quadrados a 3 variáveis obtemos não uma reta, mas sim um plano, porque o modelo de regressão linear passa a ser da forma $z = ax + by + c$, em que os números reais a, b e c são os parâmetros a determinar.

DESCRIÇÃO DO MÉTODO DOS MÍNIMOS QUADRADOS APLICADO A 3 VARIÁVEIS

Dado um conjunto de n pontos ,$n \in \mathbb{N}$, do espaço $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$, o modelo linear obtido aplicando o método dos mínimos quadrados será o plano de equação $z = ax + by + c$ que minimiza as somas dos quadrados das distâncias de cada um dos pontos ao plano.

O algoritmo tem várias etapas e envolve muitos cálculos, mas com o auxílio da calculadora conseguimos simplificar a tarefa. Com esta é possível realizar todos os somatórios pretendidos. O teu professor deve ajudar-te a perceber como realizar somatórios utilizando a calculadora.

Seja $x = \{x_1, \dots, x_n\}$.

Seja $y = \{y_1, \dots, y_n\}$.

Seja $z = \{z_1, \dots, z_n\}$.

As etapas para encontrar o modelo linear a 3 variáveis são as seguintes:

1. Calcular a média da variável x ($\bar{x}$);
2. Calcular a média da variável y ($\bar{y}$);
3. Calcular a média da variável z ($\bar{z}$);
4. Calcular $\sum_{i=1}^{n} {x_i}^2$;
5. Calcular $\sum_{i=1}^{n} {y_i}^2$;
6. Calcular $\sum_{i=1}^{n} {z_i}^2$;
7. Calcular $\sum_{i=1}^{n} x_i\, y_i$;
8. Calcular $\sum_{i=1}^{n} x_i\, z_i$;
9. Calcular $\sum_{i=1}^{n} y_i\, z_i$;
10. Calcular $SS_{xx} = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$;
11. Calcular $SS_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2$;
12. Calcular $SS_{zz} = \sum_{i=1}^{n} {z_i}^2 - n\bar{z}^2$;
13. Calcular $SS_{xy} = SS_{yx} = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\,\bar{y}$;
14. Calcular $SS_{xz} = \sum_{i=1}^{n} x_i z_i - \bar{z}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} z_i + n\bar{x}\,\bar{z}$;
15. Calcular $SS_{yz} = \sum_{i=1}^{n} y_i z_i - \bar{z}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} z_i + n\bar{y}\,\bar{z}$;

O plano é dado pela solução do sistema:

$$\begin{cases} a\, SS_{xx} + b\, SS_{xy} = SS_{xz} \\ a\, SS_{yx} + b\, SS_{yy} = SS_{yz} \\ c = \bar{z} - a\bar{x} - b\bar{y} \end{cases}$$

1.4.1. MÉTODO DOS MÍNIMOS QUADRADOS APLICADO A 3 VARIÁVEIS: EXERCÍCIO PRÁTICO

Um grande hipermercado Português implementou um novo sistema de fila de espera para o pagamento das compras dos seus clientes, o sistema de fila única, em que existe apenas uma fila e cada cliente tem de aguardar pela sua vez, posteriormente será encaminhado para uma das caixas que esteja livre.

O gerente do hipermercado queria perceber como é que o tempo médio de espera na fila, depende do número de pessoas que estão nesta e o número de caixas abertas nesse momento. Para isso considerou a variável tempo médio de espera como variável dependente das variáveis independentes número de clientes na fila e número de caixas abertas no momento.

Os dados apresentados no quadro seguinte representam o tempo de espera na fila do hipermercado, em minutos, cronometrados a 10 clientes em alturas distintas de um mesmo dia, o número de pessoas na fila e o número de caixas abertas nesses mesmos momentos do dia.

Cliente	Número de pessoas na fila (variável X)	Número de caixas abertas (variável Y)	Tempo de espera na fila em minutos (variável Z)
1	6	3	9
2	5	2	6
3	3	2	4
4	1	1	3
5	4	1	3
6	3	3	5
7	6	3	8
8	2	1	2
9	4	2	7
10	2	2	4

Tabela 3- Tabela de observações do número caixas abertas, do número de pessoas na fila e do tempo de espera na fila de um hipermercado.

Pretende-se determinar se o tempo médio de espera na fila pode ser medido em função das variáveis independentes, número de clientes na fila e número de caixas abertas no momento, através de um modelo linear.

Exercícios:

1. Com o auxílio da calculadora aplica as etapas descritas anteriormente no método dos mínimos quadrados aplicado a 3 variáveis, indicando os valores obtidos em cada etapa. No final deves indicar o modelo na forma $z = ax + by + c$.
2. Para os valores das variáveis x (número de pessoas na fila) e y (número de caixas abertas) da tabela, calcula os valores da variável z (tempo de espera na fila em minutos) dada pelo modelo linear com um arredondamento de duas casas decimais. Representa os dados numa tabela semelhante á apresentada acima, adicionado uma coluna onde terá os valores tempo de espera na fila em minutos estimado pelo modelo de regressão linear.
3. Calcula, arredondando as décimas, os valores de variável z obtido pelo modelo para os seguintes pares (x, y), $(0,1)$, $(2,3)$, $(2,4)$. Comenta os resultados obtidos de acordo com o problema. Será o modelo linear obtido um bom modelo para o problema? Será que se podia fazer algo diferente para melhorar o modelo? Faz um comentário geral do problema e da sua resolução, atenta nos valores da tabela para a chegares às tuas conclusões.

1.4.2. MÉTODO DOS MÍNIMOS QUADRADOS APLICADO A 3 VARIÁVEIS: EXERCÍCIO PRÁTICO - RESOLUÇÃO

1. Com o auxílio da calculadora aplica as etapas descritas anteriormente no método dos mínimos quadrados aplicado a 3 variáveis, indicando os valores obtidos em cada etapa. No final deves indicar o modelo na forma $z = ax + by + c$.

Aplicação do modelo linear a 3 variáveis:

1. Calcular a média da variável x ($\bar{x}$);

$$\bar{x} = \frac{18}{5}$$

2. Calcular a média da variável y ($\bar{y}$);

$$\bar{y} = 2$$

3. Calcular a média da variável z ($\bar{z}$);

$$\bar{z} = \frac{51}{10}$$

4. Calcular $\sum_{i=1}^{n} {x_i}^2$;

$$\sum_{i=1}^{n} {x_i}^2 = 156$$

5. Calcular $\sum_{i=1}^{n} {y_i}^2$;

$$\sum_{i=1}^{n} {y_i}^2 = 46$$

6. Calcular $\sum_{i=1}^{n} {z_i}^2$;

$$\sum_{i=1}^{n} {z_i}^2 = 309$$

7. Calcular $\sum_{i=1}^{n} x_i\, y_i$;

$$\sum_{i=1}^{n} x_i\, y_i = 80$$

8. Calcular $\sum_{i=1}^{n} x_i\, z_i$;

$$\sum_{i=1}^{n} x_i\, z_i = 214$$

9. Calcular $\sum_{i=1}^{n} y_i\, z_i$;

$$\sum_{i=1}^{n} y_i\, z_i = 39$$

10. Calcular $SS_{xx} = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$;

$$\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2 = \frac{132}{5}$$

11. Calcular $SS_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2$;

$$SS_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2 = 6$$

12. Calcular $SS_{zz} = \sum_{i=1}^{n} {z_i}^2 - n\bar{z}^2$;

$$SS_{zz} = \sum_{i=1}^{n} {z_i}^2 - n\bar{z}^2 = \frac{189}{10}$$

13. Calcular $SS_{xy} = SS_{yx} = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\,\bar{y}$;

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\,\bar{y} = 7$$

14. Calcular $SS_{xz} = \sum_{i=1}^{n} x_i z_i - \bar{z}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} z_i + n\bar{x}\,\bar{z}$;

$$SS_{xz} = \sum_{i=1}^{n} x_i z_i - \bar{z}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} z_i + n\bar{x}\,\bar{z} = \frac{152}{5}$$

15. Calcular $SS_{yz} = \sum_{i=1}^{n} y_i z_i - \bar{z}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} z_i + n\bar{y}\,\bar{z}$;

$$SS_{yz} = \sum_{i=1}^{n} y_i z_i - \bar{z}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} z_i + n\bar{y}\,\bar{z} = 14$$

Resolução do sistema

$$\begin{cases} a\,SS_{xx} + b\,SS_{xy} = SS_{xz} \\ a\,SS_{yx} + b\,SS_{yy} = SS_{yz} \\ c = \bar{z} - a\bar{x} - b\bar{y} \end{cases} \Leftrightarrow \begin{cases} \frac{132}{5}a + 7b = \frac{152}{5} \\ 7a + 6b = 14 \\ c = \frac{51}{10} - \frac{18}{5}a - 2b \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} 132a + 35b = 152 \\ a = \frac{14 - 6b}{7} \\ c = \frac{51}{10} - \frac{18}{5}a - 2b \end{cases} \Leftrightarrow \begin{cases} 132\left(\frac{14 - 6b}{7}\right) + 35b = 152 \\ a = \frac{14 - 6b}{7} \\ c = \frac{51}{10} - \frac{18}{5}a - 2b \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} 1848 - 792b + 245b = 1064 \\ a = \frac{14 - 6b}{7} \\ c = \frac{51}{10} - \frac{18}{5}a - 2b \end{cases} \Leftrightarrow \begin{cases} 547b = 784 \\ a = \frac{14 - 6b}{7} \\ c = \frac{51}{10} - \frac{18}{5}a - 2b \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} b = \frac{784}{547} \\ a = \frac{122}{547} \\ c = \frac{1565}{1094} \end{cases}$$

O modelo de regressão linear é dado por:

$$z = \frac{122}{547}x + \frac{784}{547}y + \frac{1565}{1094} = \frac{244x + 1568y + 1565}{1094}$$

2. Para os valores das variáveis x (número de pessoas na fila) e y (número de caixas abertas) da tabela, calcula os valores da variável z (tempo de espera na fila em minutos) dada pelo modelo linear com um arredondamento de duas casas decimais. Representa os dados numa tabela semelhante á apresentada acima, adicionado uma coluna onde terá os valores tempo de espera na fila em minutos estimado pelo modelo de regressão linear.

Cliente	Número de pessoas na fila (variável X)	Número de caixas abertas (variável Y)	Tempo de espera na fila em minutos (variável Z)	Tempo de espera na fila em minutos estimado pelo modelo de regressão linear
1	6	3	9	7,07
2	5	2	6	5,63
3	3	2	4	4,97
4	1	1	3	3,09
5	4	1	3	3,76
6	3	3	5	6,40
7	6	3	8	7,07
8	2	1	2	3,30
9	4	2	7	5,19
10	2	2	4	4,74

Tabela 4- Dados da Tabela 3 e estimativa do tempo de espera segundo o modelo linear múltiplo obtido.

3. Calcula, arredondando as décimas, os valores de variável z obtido pelo modelo para os seguintes pares (x, y), $(0,1)$, $(2,3)$, $(2,4)$. Comenta os resultados obtidos de acordo com o problema. Será o modelo linear obtido um bom modelo para o problema? Será que se podia fazer algo diferente para melhorar o modelo? Faz

um comentário geral do problema e da sua resolução, atenta nos valores da tabela para chegares às tuas conclusões.

$$z(x,y) = \frac{122}{547}x + \frac{784}{547}y + \frac{1565}{1094} = \frac{244x + 1568y + 1565}{1094}$$

$$z(0,1) = \frac{244 \times 0 + 1568 \times 1 + 1565}{1094} = \frac{3133}{1094} \approx 2{,}9min$$

$$z(2,3) = \frac{244 \times 2 + 1568 \times 3 + 1565}{1094} = \frac{6757}{1094} \approx 6{,}2min$$

$$z(2,4) = \frac{244 \times 2 + 1568 \times 4 + 1565}{1094} = \frac{8325}{1094} \approx 7{,}6min$$

Ao calcularmos $z(0,1)$ obtemos um tempo de espera de $2{,}9min$, podemos pensar neste tempo de espera como uma estimativa do tempo que cada cliente demora a pagar as suas compras. Isto porque está apenas uma caixa de atendimento a trabalhar e nenhuma pessoa na fila, concluímos que está apenas uma pessoa a pagar as suas compras.

Com o cálculo de $z(2,3)$ e $z(2,4)$ percebemos uma contradição do modelo, pois para o mesmo número de pessoas na fila, 2 pessoas, com 3 caixas de atendimento obtemos um tempo de espera de $6{,}2min$ e com 4 caixas de atendimento obtemos um tempo de espera de $7{,}6min$. Seria de esperar que se estiverem o mesmo número de pessoas numa fila mais caixas de atendimento a trabalhar deveríamos obter um tempo de espera menor. Isto explica-se, pois, nos dados obtidos na tabela resultantes da análise de alguns clientes, obtemos essa mesma contradição, para o mesmo número de pessoas na fila, 3 pessoas, obtemos um tempo de espera maior com 3 caixas abertas do que com 2.

O modelo linear é adequado para certos casos. Percebemos que para o mesmo número de caixas de atendimento a trabalhar, se aumentarmos o número de pessoas na fila aumentamos o tempo de espera na fila, o que seria de esperar. Por outro lado, com o modelo chegamos à contradição já referida, isto acontece porque talvez os clientes analisados não são os melhores para elaborar o modelo. Para chegarmos a um modelo mais eficaz deveríamos analisar mais clientes e eliminar alguns casos que se afastem do que é costume acontecer.

CAPÍTULO II- CIENTÍFICA: REGRESSÃO

2.1. REGRESSÃO LINEAR

2.1.1. MOTIVAÇÃO

O senhor Joaquim, dono da frutaria, tem uma vida muito preenchida por isso não consegue manter sempre o mesmo horário de trabalho. Assim, queria perceber como o número de horas semanais que mantinha a sua frutaria aberta afeta a sua faturação. Para isso, registou ao longo de 10 semanas, o número de horas semanais que mantinha o seu estabelecimento aberto e a respetiva faturação.

Semana	1	2	3	4	5	6	7	8	9	10
Número de horas de trabalho semanais (horas h)	50	46	54	52,5	48	57	54	48,5	55	49
Faturação semanal (euros €)	621	599	652	633	607	667	640	603	665	600

Tabela 5- Tabela de observações do número horas de trabalho semanais e a respetiva faturação semanal na frutaria do senhor Joaquim.

Se designarmos a variável número de horas semanais em que a frutaria está aberta por x e a variável faturação semanal por y, queremos estudar a forma como se relacionam as duas variáveis. Ou seja, como a variável y(faturação semanal) depende da variável x (número de horas semanais que a frutaria está aberta).

Designamos y por variável dependente e x por variável independente. Este relacionamento pode ser representado por um modelo que associa a variável independente á variável dependente. O modelo é designado por modelo de regressão linear simples, se define uma relação linear entre as variáveis, ou seja, se existe uma reta que represente bem as observações.

Vamos perceber como o modelo de regressão linear simples funciona no caso geral e para isso consideremos os seguintes pares de observações:

$$(x_1\ , y_1), (x_2\ , y_2), \dots, (x_i\ , y_i), \dots, (x_n\ , y_n)$$

Para avaliar se existe uma relação linear entre as variáveis, constrói-se o diagrama de dispersão com o conjunto de n pontos do plano. Este deve exibir uma tendência linear para que se possa usar a regressão linear, ou seja, os pontos têm de estar agrupados em redor de uma reta imaginária. Assim, a representação do diagrama de dispersão permite determinar empiricamente se existe um relacionamento linear entre as variáveis x e y, deste modo decidir se será correto aplicar o modelo de regressão linear simples.

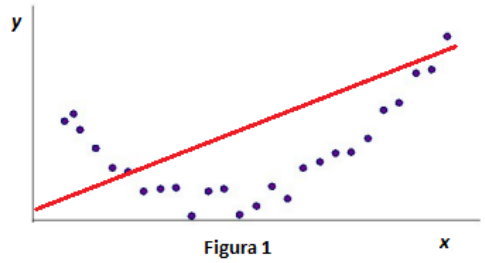


Figura 1

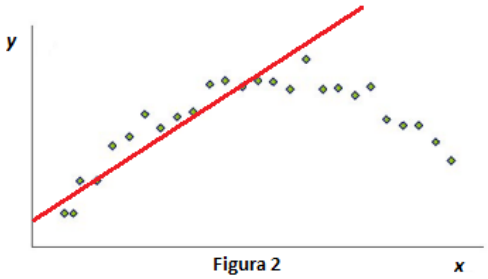


Figura 2

Os diagramas de dispersão representados pelas figuras 1 e 2 sugerem uma relação não linear, pois o conjunto de pontos não fica bem representado por uma reta.

Com a construção do diagrama de dispersão podemos também analisar a correlação linear entre as variáveis, que consiste em medir o grau de relacionamento linear entre as duas variáveis. Conseguimos também concluir empiricamente se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como os pontos se situam em redor da tal reta imaginaria que passa através do enxame de pontos. A correlação é tanto maior quanto mais próximos estão os pontos, com pequenos desvios, em relação a essa reta.

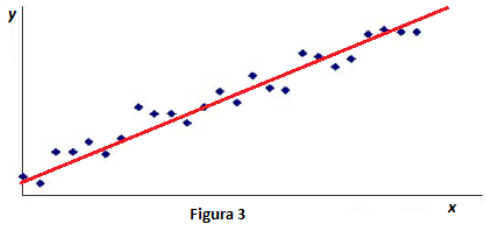


Figura 3

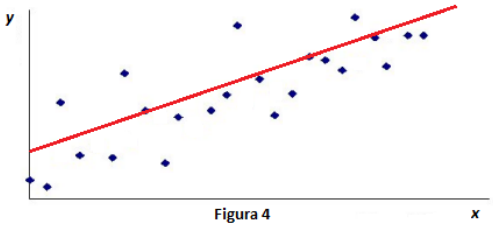


Figura 4

Os diagramas de dispersão representados pelas figuras 3 e 4 sugerem uma relação linear, os conjuntos de pontos ficam bem representados por uma reta. Evidenciamos ainda que a correlação entre as variáveis x e y é maior no diagrama de dispersão da figura 3 pois é visível que os pontos aí representados estão mais próximos da reta.

Se o declive da reta imaginária for positivo, concluímos que a correlação entre x e y é positiva, os fenómenos variam no mesmo sentido, se x cresce y também cresce. Por outro lado, se o declive da reta for negativo, então a correlação entre x e y é negativa, os fenómenos variam em sentido contrários, se x cresce y decresce.

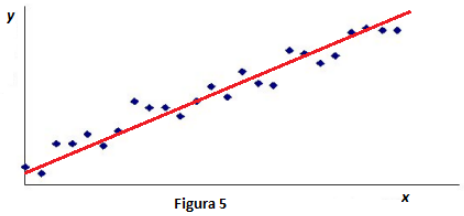


Figura 5

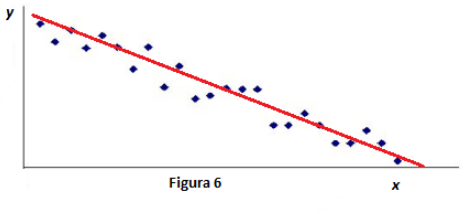
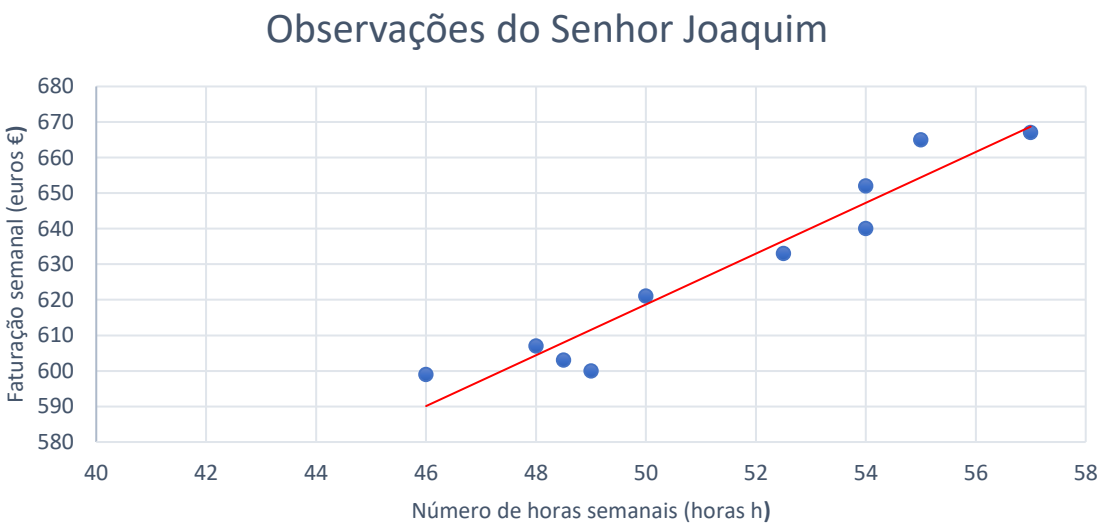


Figura 6

Os diagramas de dispersão representados pelas figuras 5 e 6 sugerem também a existência de uma relação linear, os conjuntos de pontos ficam bem representados por uma reta. O diagrama de dispersão da figura 5 sugere uma correlação positiva entre as variáveis x e y, se uma cresce a outra cresce, já o diagrama de dispersão da figura 6 sugere uma correlação negativa entre as variáveis x e y, se uma cresce a outra decresce.



Com a construção do diagrama de dispersão, utilizando os dados da tabela, notasse uma tendência linear, o conjunto de pontos fica bem representado pela reta a vermelho. Podemos ainda concluir que existe uma correlação forte e positiva entre as variáveis, visto que os pontos se localizam relativamente próximos da reta, nota-se também que se o número de horas semanais aumenta o mesmo acontece com a faturação.

2.1.2. O MODELO

O objetivo é arranjar uma única reta que aproxima o conjunto de n pontos do plano.

Considere-se o conjunto de n pontos do plano:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$$

O modelo de regressão linear é dado por:

$$y = ax + b$$

$x \rightarrow variável\ independente, sem\ erro$

$a, b \rightarrow valores\ estimados\ pelo\ modelo$

$y \rightarrow variável\ dependente$

$y_i \rightarrow obervação\ i\ da\ variável\ y$

$\hat{y}_i \rightarrow valor\ de\ y_i\ estimado\ \ pelo\ modelo\ linear$

$\varepsilon_i = y_i - \hat{y}_i$

Vamos analisar três possíveis estratégias para encontrar a reta de regressão linear.

1ª Estratégia

A primeira estratégia para encontrar a reta será minimizar a soma dos valores dos erros ε_i.

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} y_i - ax_i - b$$

Esta estratégia não funciona, como se evidencia na figura 7, em que estão representados apenas dois pontos, qualquer reta que passe no ponto médio servirá para minimizar a soma dos erros ε_i, visto que os erros cancelam.

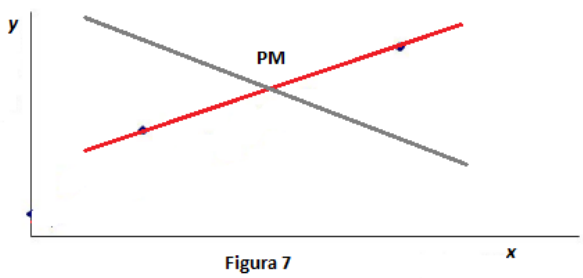


Figura 7

2ª Estratégia

A estratégia passa agora por minimizar a soma dos valores absolutos dos erros ($|\varepsilon_i|$).

$$\sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - ax_i - b|$$

Esta estratégia também não funciona. Como se verifica na figura 8, em que estão representados 4 pontos (em que cada dois têm a mesma abcissa), com estes pontos conseguimos obter duas retas, a reta a vermelho e a reta a cinzento, que minimizam a soma dos valores absolutos dos erros ($|\varepsilon_i|$). Como obtemos duas retas, esta estratégia não irá funcionar porque o objetivo é encontrar uma única reta.

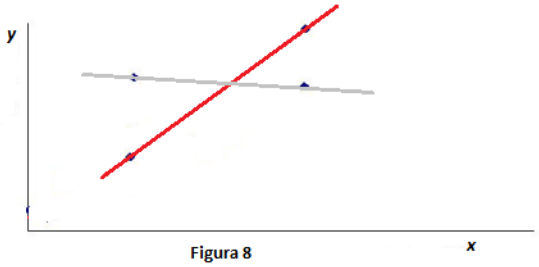


Figura 8

3ª Estratégia

Uma última estratégia que supera as deficiências das abordagens anteriores, será minimizar a soma dos quadrados dos erros ($\varepsilon_i{}^2$).

$$\sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Com esta obtemos uma única reta que relaciona as variáveis linearmente.

2.1.3. DETERMINAÇÃO DOS PARÂMETROS a E b

O objetivo passa agora por minimizar $\sum_{i=1}^{n}(y_i - ax_i - b)^2$, podemos entender este somatório como uma função S que depende das variáveis a e b, $S(a, b)$. Minimizar a função S é encontrar os seus mínimos, para achar os seus mínimos temos de calcular as derivadas parciais da função e igualá-las a zero. Mais a frente este método será explicado e demonstrado detalhadamente quando se abordar a regressão linear múltipla (Secção 2.2.).

$$(Sistema\ 1)\ \begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases}$$

Resolvendo o sistema, encontramos os valores de a e b.

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - ax_i x_i - bx_i = 0 \\ \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} {x_i}^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} 1 = 0 \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} {x_i}^2 - b\sum_{i=1}^{n} x_i = 0 \\ b = \dfrac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} {x_i}^2 - \left(\dfrac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i = 0 \\ b = \bar{y} - a\bar{x} \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} {x_i}^2 - \frac{\sum_{i=1}^{n} y_i x_i}{n} + \frac{a\sum_{i=1}^{n} {x_i}^2}{n} = 0 \\ b = \bar{y} - a\bar{x} \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} {x_i}^2 - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} + \frac{a\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n} = 0 \\ b = \bar{y} - a\bar{x} \end{cases} \Leftrightarrow$$

$$\begin{cases} a\left(\sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right) = \sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} = 0 \\ b = \bar{y} - a\bar{x} \end{cases} \Leftrightarrow$$

$$\begin{cases} a = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} {x_i}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2\right)} \\ b = \bar{y} - a\bar{x} \end{cases}$$

Os coeficientes a e b do modelo são dados por:

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}$$

$$b = \bar{y} - a\bar{x}$$

Obtemos o modelo linear simples de equação:

$$y = ax + b$$

Vamos aplicar o método dos mínimos quadrados às variáveis x (número de horas de trabalho semanais) e y (faturação semanal) da tabela 5.

$$\sum_{i=1}^{10} x_i y_i = 323980$$

$$\sum_{i=1}^{10} {x_i}^2 = 26535{,}5$$

$$\bar{x} = 51{,}4$$

$$\bar{y} = 628{,}7$$

$$a = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{10} {x_i}^2 - n\bar{x}^2} = \frac{323980 - 10 \times 51{,}4 \times 628{,}7}{26535{,}5 - 10 \times 51{,}4^2} = \frac{828{,}2}{115{,}9} \approx 7{,}15$$

$b = \bar{y} - a\bar{x} = 628{,}7 - 7{,}1458 \times 51{,}4 \approx 261{,}41$

Com os dados da tabela 5 obtemos o modelo linear simples de equação:

$$y = 7{,}15x + 261{,}41$$

Vamos agora aplicar o modelo linear simples obtido a variável x (número de horas de trabalho) da tabela 5 para obter uma estimativa da variável y (faturação semanal). Com estes dados contruímos a tabela 6.

Semana	1	2	3	4	5	6	7	8	9	10
Número de horas de trabalho semanais (horas h) Variável x	50	46	54	52,5	48	57	54	48,5	55	49
Faturação semanal (euros €) Variável y	621	599	652	633	607	667	640	603	665	600
Estimativa da faturação semanal $\hat{y}_i$(euros €)	618,91	590,31	647,51	636,79	604,61	668,96	647,51	608,19	654,66	611,76
Resíduo $\varepsilon_i = y_i - \hat{y}_i$	2,09	8,69	4,49	-3,79	2,39	-1,96	-7,51	-5,19	10,44	-11,76

Tabela 6- Dados da Tabela 5, a estimativa da faturação semanal na frutaria do Senhor Joaquim utilizando o modelo linear simples e os resíduos do modelo.

2.1.4. QUALIDADE DO AJUSTAMENTO

O modelo de regressão linear obtido pode ser visto como uma tentativa para explicar como as variações na variável dependente y resultam das alterações da variável independente x.

Uma medida útil associada ao modelo de regressão linear é comparar se as predições baseadas na reta de regressão linear superam as predições baseadas no valor médio $\bar{y}$. Para isso, analisemos os 3 tipos de variações que encontrámos no sistema.

Variação total *(SST)*

Podemos definir a variação em torno de $\bar{y}$, dar o nome de variação total e denominar por SST como:

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

Variação não explicada pelo modelo *(SSE)*

Podemos também definir a variação em torno de $\bar{y}$, dar o nome de variação em torno da reta de regressão e denominar por SSE como:

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

Variação explicada pelo modelo *(SSR)*

Podemos também definir a variação do modelo de regressão em relação a $\bar{y}$ e denominar por SSR como:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Podemos concluir que a variação total será a soma da variação que o modelo de regressão não consegue explicar mais a variação explicada pelo modelo de regressão. Ou seja,

$$SST = SSE + SSR$$

Demonstração:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 =$$

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

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

$$= SSE + SSR + \sum_{i=1}^{n}2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Temos agora de mostrar que:

$$\sum_{i=1}^{n}2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

Como $\hat{y}_i = ax_i + b$, $i \in \{1, \dots, n\}$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - ax_i - b)(ax_i + b - \bar{y}) =$$

$$= \sum_{i=1}^{n}(y_i - ax_i - b)[ax_i + (b - \bar{y})] = \sum_{i=1}^{n}(y_i - ax_i - b)\,ax_i + (y_i - ax_i - b)(b - \bar{y}) =$$

$$= a\sum_{i=1}^{n}(y_i - ax_i - b)\,x_i + (b - \bar{y})\sum_{i=1}^{n}(y_i - ax_i - b)$$

Observando que no método dos mínimos quadrados, quando resolvemos o $Sistema\ 1$ obtemos na primeira etapa:

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases}$$

Voltando a

$$a\sum_{i=1}^{n}(y_i - ax_i - b)\,x_i + (b - \bar{y})\sum_{i=1}^{n}(y_i - ax_i - b) = a \times 0 + (b - \bar{y}) \times 0 = 0$$

Logo temos que

$$SST = SSE + SSR + 0 = SSE + SSR \quad c.q.m.$$

Um critério útil para decidir se o modelo linear calculado é melhor do que utilizar simplesmente $\bar{y}$, é analisar os erros associado à reta de regressão (SSE) e o erro associado a $\bar{y}$ (SST). Se a variação não explicada pelo modelo de regressão, o erro associado à reta, for muito menor que a variação total, o erro associado a $\bar{y}$, podemos concluir que as previsões baseadas na reta serão melhores que as baseadas em $\bar{y}$.

O ajustamento será tanto melhor quanto mais pequeno for SSE em relação a SST.

Vamos analisar as variações das observações do senhor Joaquim.

Variação total: $SST = \sum_{i=1}^{10}(y_i - \bar{y})^2 = 6370{,}10$

Variação não explicada pelo modelo: $SSE = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = 454{,}59$

Variação explicada pelo modelo: $SSR = \sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 = 5915{,}51$

Analisando as variações, percebemos que a variação não explicada pelo modelo de regressão linear é muito menor do que a variação total, podemos assim concluir que as previsões do modelo se enquadram bem.

Coeficiente de determinação (r^2)

O quociente entre SSR e SST dá-nos uma medida da quantidade de variação total que pode ser explicada pelo modelo de regressão linear e representa-se por r^2.

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = 1 - \frac{SSE}{SST}$$

$r^2 \cong 0$ (valor próximo de 0) significa que grande parte da variação da variável y não pode ser explicada linearmente pela variável independente x, o modelo de regressão linear não se ajusta.

$r^2 \cong 1$ (valor próximo de 1) significa que grande parte da variação da variável y pode ser explicada linearmente pela variável independente x, fica bem ajustado pelo modelo de regressão linear.

Esta propriedade do coeficiente de determinação é explicada facilmente pela análise da fórmula do seu cálculo, pois se grande parte da variação não for explicada pelo modelo de regressão linear isto indica que $SSR \cong 0$ o que implica que $r^2 = \frac{SSR}{SST} \cong 0$, ou seja, o modelo linear não se enquadra. Se por outo lado grande parte da variação for explicada pelo modelo de regressão linear isto indica que $SSE \cong 0$ o que implica que $r^2 = 1 - \frac{SSE}{SST} \cong 1$.

Nas observações do Senhor Joaquim obtemos um coeficiente de determinação de $r^2 = \frac{SSR}{SST} = \frac{5915{,}51}{6370{,}10} \approx 0{,}93$. Este valor é relativamente próximo de 1, mais uma vez concluímos que as observações do Senhor Joaquim ficam bem ajustadas pelo modelo de regressão linear.

Coeficiente de correlação (r)

Á raiz quadrada do coeficiente de determinação dá-se o nome de coeficiente de correlação e representa-se por r.

Este coeficiente é uma medida do grau de associação linear entre o modelo de regressão linear e o conjunto de variáveis $x_1, x_2, \dots, x_n$.

$$r = \sqrt{r^2} = \sqrt{\frac{SSR}{SST}} = \sqrt{\frac{\sum_{i=1}^{n}(y - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Definindo SS_x como sendo a variação da variável x relativamente à sua média ($\bar{x}$) e calculado por:

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Definindo SS_y como sendo a variação da variável y relativamente à sua média ($\bar{y}$), já anteriormente denominado por variação total e designado por SST.

$$SST = SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Temos que:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{SS_y}} = \sqrt{\frac{\sum_{i=1}^{n}(ax_i + b - \bar{y})^2}{SS_y}} =$$

$$r = \sqrt{\frac{\sum_{i=1}^{n}(ax_i + \bar{y} - a\bar{x} - \bar{y})^2}{SS_y}} = \sqrt{\frac{\sum_{i=1}^{n}(a(x_i - \bar{x}))^2}{SS_y}} =$$

$$r = \sqrt{\frac{a^2 \sum_{i=1}^{n}((x_i - x))^2}{SS_y}} = a\sqrt{\frac{SS_x}{SS_y}}$$

Sendo a o declive da reta dada pelo modelo de regressão linear concluímos que o sinal do coeficiente de correlação será o mesmo de a pois $\sqrt{\frac{SS_x}{SS_y}} \geq 0$.

Podemos fazer assim uma análise semelhante à já feita para o coeficiente de determinação, percebendo ainda, se a correlação entre as variáveis x e y é positiva ou negativa.

$r \cong -1$ (valor próximo de 1) significa que o modelo de regressão linear se adequa, existindo uma correlação negativa entre as variáveis x e y, se uma das variáveis aumenta a outra diminuí.

$r \cong 0$ (valor próximo de 0) significa que o modelo de regressão linear não se adequa, não existindo uma correlação entre as variáveis x e y.

$r \cong 1$ (valor próximo de 1) significa que o modelo de regressão linear se adequa, existindo uma correlação positiva entre as variáveis x e y, se uma das variáveis aumenta a outra também irá aumentar.

Concluímos assim que o modelo de regressão linear se ajusta às variáveis x e y, quanto mais próximo o valor do coeficiente de correlação estiver de -1 ou 1.

Nas observações do Senhor Joaquim obtemos um coeficiente de correlação de $r \approx 0{,}96$. Este valor é ainda próximo de 1 do que o valor do coeficiente de determinação. Para além de concluirmos que as observações do Senhor Joaquim ficam bem ajustadas pelo modelo de regressão linear, concluímos também que existe uma correlação positiva entre as variáveis.

2.2. REGRESSÃO LINEAR MÚLTIPLA

2.2.1. MOTIVAÇÃO

A frutaria Senhor Joaquim tem um horário fixo de 58 horas semanais, mas como já vimos este dificilmente consegue cumpri-lo, chegando muitas vezes atrasado e por vezes saindo mais cedo.

Este, para melhor perceber se seriam os atrasos ou as saídas que afetam a faturação da frutaria, decidiu verificar nas mesmas 10 semanas qual o total do tempo dos atrasos, qual o total do tempo das saídas mais cedo do que o previsto e a respetiva faturação da semana.

Semana	1	2	3	4	5	6	7	8	9	10
Número de horas de trabalho semanais (horas h)	50	46	54	52,5	48	57	54	48,5	55	49
Tempo semanal total de atrasos (minutos m)	268	420	60	134	356	34	52	298	32	248
Tempo semanal total de saídas antes do previsto (minutos m)	212	300	180	196	244	26	188	272	148	292
Faturação semanal (euros €)	621	599	652	633	607	667	640	603	665	600

Tabela 7- Tabela de observações do número horas de trabalho semanais, tempo semanal total de atrasos, o tempo semanal de saídas antes do previsto e a respetiva faturação semanal na frutaria do senhor Joaquim.

Se designarmos a variável tempo semanal total de atrasos por x_1,a variável tempo semanal total de saídas antes do previsto por x_2 e a variável faturação semanal por y, queremos estudar a forma como se relacionam as três variáveis. Ou seja, como a variável y (faturação semanal) depende das variáveis x_1 (tempo semanal total de atraso) e x_2 (tempo semanal total de saídas).

Designamos y por variável dependente, x_1 e x_2 por variáveis independentes. Este relacionamento pode ser representado por um modelo que associa as variáveis independentes á variável dependente. O modelo é designado por modelo de regressão linear simples se define uma relação linear entre as variáveis, ou seja, se existe neste caso um plano que representa bem as observações. Um plano porque agora como temos observações de 3 variáveis passamos para $\mathbb{R}^3$.

2.2.2. O MODELO

O modelo de regressão linear múltipla é uma extensão do modelo de regressão linear simples.

O objetivo é arranjar um modelo linear que melhor relaciona uma variável dependente y com um conjunto de j variáveis independentes, $(x_1, \dots , x_j)$, em que temos n observações dessas variáveis.

No fundo temos n pontos de $\mathbb{R}^{j+1}$

$$\{(x_{11}, \dots, x_{j1}, y_1); (x_{12}, \dots, x_{j2}, y_2); \dots; (x_{1n}, \dots , x_{jn}, y_n)\}$$

$x_{im} \rightarrow$ $observação\ m\ da\ variável\ independente\ x_i$

O modelo de regressão linear múltipla:

$$y = a_1x_1 + a_2x_2 + \cdots + a_jx_j + b$$

$x_i \rightarrow variável\ independente$

$\hat{y}_i \rightarrow valor\ estimado\ de\ y_i\ pelo\ modelo\ linear$

$\varepsilon_i = y_i - \hat{y}_i$

$a_1, \dots, a_j, b \rightarrow valores\ estimados\ pelo\ modelo$

$y \rightarrow variável\ independente$

De forma semelhante ao que acontece no método dos mínimos quadrados para duas variáveis, sendo este uma generalização, a estratégia utilizada para encontrar os parâmetros $a_1, \dots, a_j, b$ é a mesma, minimizar o quadrado dos desvios, ε_i.

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b)^2$$

2.2.3. DETERMINAÇÃO DOS PARÂMETROS $a_1, \dots, a_j, b$

O objetivo é minimizar $\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b)^2$ que pode ser visto como uma função S que depende das variáveis $a_1, \dots, a_j, b$, $S(a_1, \dots, a_j, b)$. Minimizar a função S é o mesmo que encontrar os seus mínimos, para achar os seus mínimos temos de calcular as derivadas parciais e igualá-las a zero como fizemos anteriormente no método dos mínimos quadrados para duas variáveis. Este método será explicado posteriormente.

Para determinar $a_1, \dots, a_j, b$ de modo a minimizar S tem de se resolver o seguinte sistema de equações:

$$\frac{\partial S}{\partial a_1} = 0 \wedge \frac{\partial S}{\partial a_2} = 0 \wedge \dots \wedge \frac{\partial S}{\partial a_j} = 0 \wedge \frac{\partial S}{\partial b} = 0 \Leftrightarrow$$

$$\Leftrightarrow (Sistema\ 2)\begin{cases} (-2)\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b)\,x_{1i} = 0 \\ \dots \\ (-2)\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b)\,x_{ji} = 0 \\ (-2)\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b) = 0 \end{cases}$$

A última equação permite determinar o valor do estimador b, que é idêntico ao que foi definido para o modelo de regressão linear simples.

$$(-2)\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b) = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n}(y_i - a_1x_{1i} - a_2x_{2i} - \cdots - a_jx_{ji} - b) = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n}y_i - a_1\sum_{i=1}^{n}x_{1i} - a_2\sum_{i=1}^{n}x_{2i} - \cdots - a_n\sum_{i=1}^{n}x_{ji} - \sum_{i=1}^{n}b = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n}y_i - a_1\sum_{i=1}^{n}x_{1i} - a_2\sum_{i=1}^{n}x_{2i} - \cdots - a_j\sum_{i=1}^{n}x_{ji} = nb \Leftrightarrow$$

$$\Leftrightarrow b = \frac{\sum_{i=1}^{n}y_i - a_1\sum_{i=1}^{n}x_{1i} - a_2\sum_{i=1}^{n}x_{2i} - \cdots - a_j\sum_{i=1}^{n}x_{ji}}{n}$$

Obtemos:

$$b = \bar{y} - a_1\overline{x_1} - a_2\overline{x_2} - \cdots - a_j\overline{x_j}$$

Desenvolvendo as restantes equações, obtém-se o seguinte sistema cuja resolução permite obter os parâmetros $a_1, \dots, a_j$.

$$\begin{cases} a_1 S_{x_1x_1} + a_2 S_{x_1x_2} + \cdots + a_j S_{x_1x_j} = S_{x_1y} \\ a_1 S_{x_2x_1} + a_2 S_{x_2x_2} + \cdots + a_j S_{x_2x_j} = S_{x_2y} \\ \\ \cdots \\ \\ a_1 S_{x_jx_1} + a_2 S_{x_jx_2} + \cdots + a_j S_{x_jx_j} = S_{x_jy} \end{cases}$$

Onde:

$S_{x_kx_f} = \sum_{i=1}^{n}(x_{ki} - \overline{x_k})(x_{fi} - \overline{x_f})$ para $k, f \in \{1, \dots, j\}$

$S_{x_ky} = \sum_{i=1}^{n}(x_{ki} - \overline{x_k})(y_i - \bar{y})$ para $k \in \{1, \dots, j\}$

$x_{ki} \rightarrow observação\ i\ da\ variável\ x_k$

$\overline{x_k} \rightarrow média\ aritmética\ da\ variável\ x_k$

$y_i \rightarrow observação\ i\ da\ variável\ y$

$\bar{y} \rightarrow média\ aritmética\ da\ variável\ y$

Vamos aplicar o método dos mínimos quadrados às variáveis x_1 (tempo semanal total de atrasos), x_2 (tempo semanal total de saídas antes do previsto) e y (faturação semanal) aos dados da tabela 7.

$\overline{x_1} = 190{,}2$

$\overline{x_2} = 205{,}8$

$\bar{y} = 628{,}7$

$S_{x_1x_1} = \sum_{i=1}^{10}(x_{1i} - \overline{x_1})^2 = 189947{,}6$

$S_{x_1x_2} = \sum_{i=1}^{10}(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_f}) = 84180{,}4$

$S_{x_2x_2} = \sum_{i=1}^{10}(x_{2i} - \overline{x_2})^2 = 58931{,}6$

$S_{x_1y} = \sum_{i=1}^{n}(x_{1i} - \overline{x_1})(y_i - \bar{y}) = -32013{,}4$

$S_{x_2y} = \sum_{i=1}^{n}(x_{2i} - \overline{x_2})(y_i - \bar{y}) = -17678{,}6$

$$\begin{cases} a_1 S_{x_1x_1} + a_2 S_{x_1x_2} = S_{x_1y} \\ a_1 S_{x_2x_1} + a_2 S_{x_2x_2} = S_{x_2y} \end{cases} \Leftrightarrow \Leftrightarrow \begin{cases} 189947{,}6\ a_1 + 84180{,}4\ a_2 = -32013{,}4 \\ 84180{,}4\ a_1 + 58931{,}6\ a_2 = -17678{,}6 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} a_1 = -\frac{2766731}{28524835} \\ a_2 = -\frac{9209817}{57049670} \end{cases}$$

$$b = \bar{y} - a_1\overline{x_1} - a_2\overline{x_2} = 628{,}7 + \frac{2766731}{28524835} \times 190{,}2 + \frac{9209817}{57049670} \times 205{,}8 \approx 680{,}4$$

Obtemos o modelo linear múltiplo de equação:

$$y = a_1 x_1 + a_2 x_2 + b \Leftrightarrow$$

$$\Leftrightarrow y = -\frac{2766731}{28524835} x_1 + -\frac{9209817}{57049670} x_2 + 680{,}4$$

Vamos agora aplicar o modelo obtido às variáveis da tabela 7, x_1 (tempo semanal total de atrasos) e x_2 (tempo semanal total de saídas antes do previsto) para obter uma estimativa da variável y (faturação semanal). Com estes dados contruímos a tabela 8.

Semana	1	2	3	4	5	6	7	8	9	10
Tempo semanal total de atrasos x_1(minutos m)	268	420	60	134	356	34	52	298	32	248
Tempo semanal total de saídas antes do previsto x_2 (minutos m)	212	300	180	196	244	26	188	272	148	292
Faturação semanal y(euros €)	621	599	652	633	607	667	640	603	665	600
Estimativa da faturação semanal utilizando o modelo linear múltiplo $\hat{y}_i$(euros €)	620,18	591,23	645,52	635,76	606,48	672,90	645,01	607,59	653,40	609,21
Resíduo do modelo linear múltiplo $\varepsilon_i = y_i - \hat{y}_i$	0,82	7,77	6,48	-2,76	0,56	-5,90	-5,01	-4,59	11,6	-9,21

Tabela 8- Dados da Tabela 7, a estimativa da faturação semanal na frutaria do Senhor Joaquim utilizando o modelo linear múltiplo e os resíduos do modelo.

2.2.4. DETERMINAÇÃO DOS PARÂMETROS $a_1, \dots, a_j, b$ UTILIZANDO MATRIZES

Para determinar os coeficientes $a_1, \dots, a_j, b$ do modelo de regressão linear através do processo anterior temos de fazer bastantes contas, visto ser necessário calcular todos os $S_{x_k x_f}$ para todas as variáveis x_k e x_f. Mas podemos resolver o sistema inicial através de matrizes, o que torna a solução mais simples se tivermos algum meio de calcular produtos e inversas de matrizes facilmente.

Voltando ao sistema:

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_j x_{ji} - b\right) x_{1i} = 0 \\ \dots \\ \sum_{i=1}^{n} \left(y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_j x_{ji} - b\right) x_{ji} = 0 \\ \sum_{i=1}^{n} \left(y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_j x_{ji} - b\right) = 0 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i x_{1i} - a_1 x_{1i} x_{1i} - a_2 x_{2i} x_{1i} - \cdots - a_j x_{ji} x_{1i} - b x_{1i} = 0 \\ \dots \\ \sum_{i=1}^{n} y_i x_{ji} - a_1 x_{1i} x_{ji} - a_2 x_{2i} x_{ji} - \cdots - a_j x_{ji} x_{ji} - b\, x_{ji} = 0 \\ \sum_{i=1}^{n} y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_j x_{ji} - b = 0 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i x_{1i} = a_1 \sum_{i=1}^{n} x_{1i} x_{1i} + a_2 \sum_{i=1}^{n} x_{2i} x_{1i} + \cdots + a_j \sum_{i=1}^{n} x_{2i} x_{1i} + b \sum_{i=1}^{n} x_{1i} \\ \dots \\ \sum_{i=1}^{n} y_i x_{ji} = a_1 \sum_{i=1}^{n} x_{1i} x_{ji} + a_2 \sum_{i=1}^{n} x_{2i} x_{ji} + \cdots + a_j \sum_{i=1}^{n} x_{2i} x_{ji} + b \sum_{i=1}^{n} x_{ji} \\ \sum_{i=1}^{n} y_i = a_1 \sum_{i=1}^{n} x_{1i} + a_2 \sum_{i=1}^{n} x_{2i} + \cdots + a_j \sum_{i=1}^{n} x_{2i} + bn \end{cases}$$

Podemos agora escrever este sistema em forma matricial, para isso consideremos as seguintes matrizes:

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ b \end{pmatrix}$$

$$Y = \begin{pmatrix} \sum_{i=1}^{n} y_i x_{1i} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ji} \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

$$X = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}x_{1i} & \sum_{i=1}^{n} x_{2i}x_{1i} & \cdots & \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{1i} \\ \vdots & & & \vdots & \\ \sum_{i=1}^{n} x_{1i}x_{ji} & \sum_{i=1}^{n} x_{2i}x_{ji} & & \sum_{i=1}^{n} x_{2i}x_{ji} & \sum_{i=1}^{n} x_{ji} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & a_j \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} 1 \end{pmatrix}$$

Obtemos o sistema:

$$Y = X\,A$$

Convém realçar que as dimensões das matrizes estão coerentes, sendo X uma matriz quadrada com $j+1$ linhas e $j+1$ colunas e A uma matriz coluna de $j+1$ linha, temos que o produto da matriz X pela matriz A será também uma matriz coluna com $j+1$ linha, o que acontece com a matriz Y.

Através de operações com matrizes conseguimos resolver o sistema acima isolando a matriz A que nos irá dar os valores do modelo linear que queremos estimar, $a_1, \dots, a_j, b$.

$$Y = X\,A$$

Considerando:

X^T a matriz transporta de X.

X^{-1} a matriz inversa da matriz X.

I_d como a matriz identidade com d linhas e d colunas.

$(X^T)^{-1}X^T = X^T(X^T)^{-1} = I_{j+1}$

$$Y = X\,A$$

XA é uma matriz coluna com $j + 1$ linha

Podemos multiplicar por X^T dos dois lados da equação pois é uma matriz quadrada,$j + 1$ linhas e $j + 1$ colunas. Obtemos:

$$X^T\,Y = X^T\,X\,A$$

Em cada lado da equação temos matrizes colunas com j+1 linhas.

A matriz X^TX é uma matriz quadrada, com $j + 1$ linhas e $j + 1$ colunas, portanto a sua inversa, $(X^TX)^{-1}$, terá as mesmas dimensões. Podemos multiplicar a esquerda de ambos os lados da equação por $(X^TX)^{-1}$. Obtemos:

$$(X^TX)^{-1}X^T\,Y = (X^TX)^{-1}X^T\,X\,A$$

Como $(X^TX)^{-1}X^T\,X = I_{J+1}$

$$(X^TX)^{-1}X^T\,Y = I_{J+1}\,A$$

Como $I_{J+1}\,A = A$

$$(X^TX)^{-1}X^T\,Y = A$$

$$A = (X^TX)^{-1}X^TY$$

Obtém-se os coeficientes do modelo de regressão linear pelo cálculo da matriz $(X^TX)^{-1}X^TY$

$$A^* = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_j \\ b \end{pmatrix} = (X^TX)^{-1}X^TY$$

Apesar de se conseguir encontrar uma fórmula que se chegue a solução A^*, esta devido a inversão da matriz X^TX pode ser numericamente instável, isto é, pode levar a aproximações pouco exatas de A^*.

2.2.5. QUALIDADE DO AJUSTAMENTO

O estudo que se faz para a qualidade do ajustamento na regressão linear múltipla é semelhante ao já feito para a regressão linear, as variações são calculadas da mesma forma. Comparar os erros associado ao modelo de regressão (SSE) com o erro associado a $\bar{y}$ (SST), se o erro associada ao modelo for muito menor que o erro associada a $\bar{y}$, podemos concluir que as previsões baseadas no modelo serão melhores que as baseadas em $\bar{y}$.

Tal como acontece no modelo linear simples o ajustamento será tanto melhor quanto mais pequeno for SSE em relação a SST.

Vamos calcular as variações das observações do senhor Joaquim.

Variação total: $SST = \sum_{i=1}^{10}(y_i - \bar{y})^2 = 6370{,}10$

Variação não explicada pelo modelo: $SSE = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = 411{,}33$

Variação explicada pelo modelo: $SSR = \sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 = 5958{,}77$

Nota-se que $SSE = 411{,}33$ é bastante menor do que $SSR = 5958{,}77$, podemos assim concluir que o ajustamento é bastante adequado.

Coeficiente de determinação e de correlação

Estes coeficientes são definidos da mesma forma que na regressão linear**.**

O coeficiente de determinação é representado por r^2, resulta do quociente entre SSR e SST e dá-nos uma medida da quantidade de variação total que pode ser explicada pelo modelo de regressão linear.

$$r^2 = \frac{SSR}{SST}$$

À raiz quadrada do coeficiente de determinação dá-se o nome de coeficiente de correlação e representa-se por r.

$$r = \sqrt{\frac{SSR}{SST}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{SS_y}} =$$

Analisando os coeficientes das observações do senhor Joaquim, obtemos que $r^2 = \frac{SSR}{SST} = \frac{5958{,}77}{6370{,}10} \approx 0{,}9354$ e $r \approx 0{,}9671$.

Tanto o coeficiente de determinação como o de correlação estão próximos de 1, o que indica que o modelo de regressão linear múltipla se ajusta às variáveis.

2.2.6. PROCURA DO MÍNIMO DA FUNÇÃO S

Teorema 1

Se $f: \mathbb{R}^n \to \mathbb{R}$ for uma função derivável e admitir o extremo $(x_1, x_2, \dots, x_n)$ então cada derivada parcial de primeira ordem da função f anula-se para esses valores das variáveis independentes. Ou seja, $\forall\, i \in \{1, \dots, n\}\ \frac{\partial f}{\partial x_i}\Big|_{(x_1, x_2, \dots, x_n)} = 0$

Consideremos a função $S: \mathbb{R}^{j+1} \to \mathbb{R}$ definida por:

$$S(a_1, \dots, a_j, b) = \sum_{i=1}^{n} (y_i - a_1 x_{1i} - a_2 x_{2i} - \cdots - a_j x_{ji} - b)^2 =$$

$$= (y_1 - a_1 x_{11} - a_2 x_{21} - \cdots - a_j x_{j1} - b)^2 + \cdots + (y_n - a_1 x_{1n} - a_2 x_{2n} - \cdots - a_j x_{jn} - b)^2$$

Para simplificar podemos escrever a função S em notação matricial, para isso consideremos as seguintes n funções $R_i: \mathbb{R}^n \to \mathbb{R}$ definidas por:

$$R_i(a_1, \dots, a_j, b) = y_i - X_i A$$

em que A e X_i são definidos por:

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_j \\ b \end{pmatrix}$$

$$X_i = (x_{1i} \quad x_{2i} \quad \dots \quad x_{ji} \quad 1)\ \forall\, i \in \{1, \dots, n\}$$

Temos então que:

$$S(a_1, \dots, a_j, b) = \sum_{i=1}^{n} [y_i - X_i A]^2 = \sum_{i=1}^{n} \left[R_i(a_1, \dots, a_j, b)\right]^2$$

Para aplicarmos o teorema 1 á função S temos de garantir primeiro a sua derivabilidade.

Sabemos que cada uma das funções R_i é derivável em $\mathbb{R}^{j+1}$ em relação a cada uma das variáveis $a_1, \dots, a_j, b$ e que a função $f(x) = x^2\ definida\ de\ \mathbb{R}\ em\ \mathbb{R}$ também é derivável em $\mathbb{R}$. Utilizando o teorema que afirma que a composta de funções deriváveis é derivável obtemos que as funções ${R_i}^2$ são deriváveis.

A função S é obtida através da soma das n funções ${R_i}^2$. A soma de funções deriváveis resulta em uma função derivável. Concluímos assim que a função S é derivável em relação a cada uma das variáveis $a_1, \dots, a_j, b$.

Aplicando o teorema 1 á função S obtemos o sistema:

$$\frac{\partial S}{\partial a_1} = 0 \wedge \frac{\partial S}{\partial a_2} = 0 \wedge \dots \wedge \frac{\partial S}{\partial a_j} = 0 \wedge \frac{\partial S}{\partial b} = 0$$

Este sistema já foi resolvido na secção 2.2.4., onde se encontrou como solução a matriz:

$$A^* = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_j \\ b \end{pmatrix}$$

Que pode ser vista como o ponto $a^* = (a_1, a_2, \dots, a_j, b)$ que pelo teorema 1 será um candidato a ponto extremo da função S.

Temos agora de garantir que o ponto a^* se trata de um ponto mínimo da função S, para isso consideremos $h \in \mathbb{R}^{j+1}$ um vetor arbitrário não nulo definido por:

$$h = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_j \\ h_{j+1} \end{pmatrix}$$

Calculamos $S(a^* + h)$.

$$S(a^* + h) = \sum_{i=1}^{n} [R_i(a^* + h)]^2 = \sum_{i=1}^{n} [y_i - X_i(a^* + h)]^2 = \sum_{i=1}^{n} [y_i - X_i a^* - X_i\, h]^2 =$$

$$= \sum_{i=1}^{n} [R_i(a^*) - X_i\, h]^2 = \sum_{i=1}^{n} [R_i(a^*)]^2 - 2\sum_{i=1}^{n} R_i(a^*)\, X_i\, h + \sum_{i=1}^{n} (X_i\, h)^2$$

Temos que $\sum_{i=1}^{n} R_i(a^*)\, X_i = 0$, este resultado vem da última equação do sistema quando se calcula as derivadas laterais

$$\left.\frac{\partial S}{\partial b}\right|_{a^*} = 0 \Leftrightarrow \left.\frac{\partial \sum_{i=1}^{n} \left[R_i\left(a_1, \dots, a_j, b\right)\right]^2}{\partial b}\right|_{a^*} = 0 \Leftrightarrow -2\sum_{i=1}^{n} R_i(a^*)\, X_i = 0 \Leftrightarrow$$

$$\Leftrightarrow -2h\sum_{i=1}^{n} R_i(a^*)\, X_i = 0 \Leftrightarrow -2\sum_{i=1}^{n} R_i(a^*)\, X_i\, h = 0$$

Logo,

$$S(a^* + h) = \sum_{i=1}^{n} [R_i(a^*)]^2 + 0 + \sum_{i=1}^{n} (X_i\, h)^2$$

Como as matrizes X_i e h são não nulas temos que:

$$\sum_{i=1}^{n} (X_i\, h)^2 > 0$$

Assim, podemos afirmar que:

$$S(a^* + h) > \sum_{i=1}^{n} [R_i(a^*)]^2 = S(a^*)$$

Obtemos que:

$$\forall\, h \in \mathbb{R}^{j+1}, \qquad S(a^*) < S(a^* + h)$$

Estamos deste modo em condições de garantir que a^* é um ponto mínimo da função S.

CONCLUSÃO E AGRADECIMENTOS

Este ano letivo contribuiu muito para o meu desenvolvimento como futuro profissional de educação, aprendi imensas estratégias de ensino que não conhecia e mais importante de tudo desenvolvi e adquiri muitas capacidades. Foi uma experiência muito gratificante poder trabalhar como professor, esta profissão revelou ainda mais o gosto que eu tenho por ensinar e a paixão que eu tenho pela matemática.

Transmitir conhecimento de geração em geração é fundamental para a sobrevivência e o desenvolvimento das civilizações, a profissão de professor tem como principal objetivo isso mesmo. Tudo isto incute em nós professores uma sensação de sermos úteis e fundamentais para a comunidade.

Aprendi que um professor tem de ser também um educador, principalmente nas turmas do básico em que os alunos são jovens e irrequietos. Se o professor não consegue controlar os estudantes, estes irão comprometer a aula.

Gostei muito de poder estagiar na Escola João Gonçalves Zarco pois para além de contribuir imenso para a minha formação profissional foi muito importante para crescer enquanto pessoa. O contacto com os jovens estudantes, os outros professores e todos os intervenientes na escola fez-me alargar a visão que eu tinha desta, pois cada um destes tem uma função e motivação diferente para fazer parte do ambiente escolar e assim contribuir para que a escola seja um ótimo lugar para aprender.

Queria agradecer à minha família e à minha namorada que sempre me apoiaram a seguir o ramo do ensino, às professoras orientadoras Maria João Rodrigues e Maria João Costa que estavam sempre disponíveis para esclarecer qualquer dúvida que surgisse, ao meu colega Nuno Ferreira que me acompanhou da melhorar maneira e à professora Dilma Tuna que me ensinou o que realmente é ensinar.

BIBLIOGRAFIA

BARNES, Randal J.- **Matrix Differentiation (and someother stuff).** [2006]. 9 diapositivos. Acessível no Departamento de Engenharia Civil, Universidade do Minnesota, Minneapolis, USA.

CHAPRA, Steven C. e CANALE, Raymond P.- **Numerical Methods for Engineers**. 6ª Edição, Mcgraw-Hill Eduction, 1985.

CORNILLON, Pierre A. e LOBER, Eric M.- **Régression avec R**. 1ª Edição, Springer, 2010.

DAVID, Cruise- **Why is SST= SSE + SSR?,** 2019. [Acedido em: 23 abril 2020]. Disponível em WWW: <URL:https://stats.stackexchange.com/questions/207841/why-is-sst-sse-ssr-one-variable-linear-regression

KENDALL, Atinkson E.- **An introduction to numerical analysis.** 2ª Edição, John Wiley & Sons, Inc., 1989.

LOBO, Bernardo A.- **Métodos Estatísticos de Previsão**. 80 diapositivos. Disponível em WWW:
<URL:https://sigarra.up.pt/feup/pt/conteudos_service.conteudos_cont?pct_id=33808&pv_cod=03awyHjmGJp7

RODRIGUES, Maria J.- **Análise Numérica**. [2017/2018]. 129 diapositivos. Acessível na Faculdade de Ciências, Porto, Portugal.