Sumário pormenorizado da lição

Regressão linear múltipla

Andreia Teixeira Marques Dionísio Basílio

Universidade de Évora

Para candidatura ao título académico de agregado pela Universidade de Évora,
nos termos do Decreto-Lei n.º 239/2007 de 19 de junho

Fevereiro de 2023

Este sumário pormenorizado da lição "Regressão linear múltipla" constitui um dos elementos apresentados por *Andreia Teixeira Marques Dionísio Basílio* para candidatura ao título académico de agregado pela Universidade de Évora no ramo do conhecimento de Gestão, nos termos do Decreto-Lei nº 239/2007 de 19 de junho.

Índice

1. INTRODUÇÃO

Este documento será maioritariamente escrito na primeira pessoa (singular e plural). Ao contrário de outros documentos, aqui considero que há toda uma envolvência pessoal, que me impede de escrever uniformemente no impessoal.

A unidade curricular (UC) de Análise de Dados para Negócios I é em si uma disciplina que me apaixona, cujos conteúdos gosto muito de lecionar e partilhar com os estudantes. É-me, por isso, muito difícil selecionar um tópico específico para esta Lição. Desde a descrição das variáveis, a respetiva visualização e exploração de características, passando pela inferência e terminando nos modelos de regressão linear, em todas as matérias existem elementos fascinantes que tornam a análise de dados uma disciplina de grande interesse e utilidade para a gestão, nas suas diversas áreas de especialização.

Reconhecendo-se a importância, utilidade e aplicabilidade prática dos conteúdos lecionados para as diversas áreas de especialização, procurei um tópico que fosse igualmente importante para todas as áreas de especialização do Mestrado em Gestão. Tendo em consideração a duração da Lição, 50 minutos, parece-me importante a escolha de um tópico relativamente "independente", que não obrigasse a rever muitos conceitos abordados em momentos anteriores da unidade curricular e que simultaneamente proporcionasse momentos de explanação teórica, análise de dados com recurso a *software*, interpretação dos resultados e intuição económica e de gestão.

Neste contexto selecionei o tópico "*Regressão linear múltipla*". Pensar-se-á que pouco mais há a dizer sobre o mesmo, mas o meu objetivo neste documento e nesta Lição não é acrescentar algo novo, é sim mostrar e salientar as potencialidades deste modelo, especialmente no que respeita à avaliação de determinantes estatisticamente significativos de variáveis que pretendemos explicar. De referir que, na área da gestão, muitos artigos publicados incluem a regressão linear múltipla como método fundamental para obtenção de resultados e para atingir os objetivos especificados. Além disso, a regressão linear constitui a base para o desenvolvimento e compreensão de muitos métodos estatísticos e econométricos que são frequentemente encontrados em trabalhos de investigação na área de gestão, nomeadamente os modelos de equações estruturais (SEM), *partial least squares* (PLS), modelos heterocedásticos (ARCH, GARCH e famílias), entre muitos outros.

O tema desta Lição, "Regressão linear múltipla", terá um cariz teórico-prático. Numa primeira fase é feita a ligação entre a regressão linear simples e a regressão linear múltipla, no sentido de justificar a utilização de mais variáveis no modelo de regressão com vista à melhor explicação da variável dependente. O Método dos Mínimos Quadrados (MMQ), já apresentado em aula anterior aquando da regressão linear simples, é aqui apresentado, essencialmente, enquanto generalização do primeiro, sendo explorado na sua versão matricial. Os pressupostos do Teorema de Gauss-Markov são também alvo de análise, sendo apresentados de modo mais intuitivo, dado terem sido alvo de apresentação detalhada aquando da explicação do modelo de regressão linear simples. O enfoque será então dado à estimação de modelos de regressão linear múltipla, interpretação dos resultados, inferência para coeficientes individuais, para o modelo global e para a comparação entre o modelo global e um modelo restrito. Será também explorada a utilização de variáveis binárias enquanto variáveis independentes, dando especial atenção à sua construção, interpretação enquanto variáveis que podem promover alterações na constante ou alterações no declive da equação.

O documento apresentado, além desta seção introdutória, é composto por mais três secções. Na secção dois apresentam-se os objetivos pedagógicos e o plano da Lição de Agregação. Na secção três, e aquela que considero ser a principal, descrevem-se os conteúdos da Lição em si. E, na quarta e última secção deste documento/relatório apresentam-se as considerações finais.

2. OBJETIVOS E PLANO DA LIÇÃO

2.1. Objetivos pedagógicos

Nesta Lição o tema em estudo é a regressão linear múltipla, enquanto ferramenta que ajuda a encontrar os determinantes de uma determinada variável, de modo a explicar e prever o seu comportamento. O primeiro objetivo consiste em consolidar os conhecimentos já adquiridos em aula anterior acerca do modelo de regressão linear simples, mais concretamente no que se refere ao método de estimação (o método dos mínimos quadrados),

respetivos pressupostos, interpretação dos coeficientes e inferência.

Deste modo, o modelo de regressão linear múltipla é apresentado enquanto uma generalização do modelo de regressão linear simples, sendo dada espacial ênfase às suas características e potencialidades. Neste contexto, a motivação para a introdução do modelo de regressão linear múltipla, constitui o segundo objetivo desta Lição, que integra igualmente a interpretação dos respetivos coeficientes. A interpretação dos resultados obtidos assume muita importância nesta unidade curricular, uma vez que o intuito da mesma é a utilização de ferramentas que apoiam a tomada de decisão.

A compreensão da generalização do MMQ para o modelo de regressão linear múltipla, constitui o terceiro objetivo. Pretende-se que, de forma intuitiva, os estudantes compreendam a derivação dos estimadores do MMQ e respetivos pressupostos. Para além dos pressupostos do Teorema de Gauss-Markov já explorados aquando da estimação do modelo de regressão linear simples, é acrescentado o pressuposto de ausência de colinearidade perfeita. Neste contexto, o objetivo será compreender a importância dos cuidados a ter aquando da estimação deste modelo através do MMQ e a necessidade de avaliar os respetivos pressupostos. A verificação destes pressupostos não faz parte do plano desta Lição.

Um quarto objetivo prende-se com a exploração inferencial no modelo de regressão linear. Desde os testes de hipóteses para parâmetros individuais, ao teste à significância global do modelo (que serão revisitados, uma vez que já foram abordados no modelo de regressão linear simples), sendo agora acrescentado o teste para comparação entre um modelo restrito e o modelo global.

Finalmente, o quinto e último objetivo prende-se com a utilização de informação qualitativa no modelo de regressão linear múltipla, através da inclusão de variáveis binárias enquanto variáveis independentes. Pretende-se que os estudantes compreendam a construção e formas de modelação de variáveis qualitativas no modelo de regressão linear e que compreendam o impacto que uma variável binária pode ter no modelo, seja ao nível da constante, seja no declive da equação.

2.2. Plano da Lição

A presente Lição terá a duração de 50 minutos. O plano previsto é o seguinte:

1. Muito breve revisão dos conceitos referentes ao modelo de regressão linear simples

2. Motivação para o modelo de regressão linear múltipla, estimação do modelo e sua interpretação

3. Pressupostos para a regressão linear múltipla

4. Avaliação do ajustamento e inferência no modelo de regressão linear múltipla

5. Modelo de regressão linear múltipla com informação qualitativa

O primeiro ponto da Lição é incluído para que a audiência da Lição de Agregação (e não apenas o júri) possa acompanhar sem grandes problemas a Lição. Se a aula fosse com alunos de Análise de Dados para Negócios I, este ponto não seria necessário porque teria sido convenientemente explorado em aulas anteriores.

2.3. Bibliografia recomendada

A bibliografia recomendada aos alunos para esta lição é a seguinte:

- Dionísio, A. (2019). *Análise de Dados para Negócios,* Texto de Apoio para as unidades curriculares Análise de Dados para Negócios I e Análise de Dados para Negócios II do Curso de Mestrado em Gestão, Universidade de Évora, ISBN 978-972-778-136-2.
- Newbold, P., Carlson, W. e Thorne, B. (2019). *Statistics for Business and Economics*, 9th edition, Pearson Education Limited, UK.
- Wooldridge, J. (2019). *Introductory Econometrics – A Modern Approach*, 7th edition, South-Western College Publishing, Thomson Learning. Florence.

3. SÍNTESE DOS CONTEÚDOS ABORDADOS

3.1. Muito breve revisão dos conceitos dos conceitos referentes ao modelo de regressão linear simples

A análise de regressão e de correlação envolve a análise de dados amostrais para avaliar o modo como duas ou mais variáveis estão relacionadas entre si numa população. A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear que descreva o relacionamento entre duas variáveis.

Se se tiverem duas variáveis aleatórias *X* e Y, em que se considera que *Y* é a variável explicada (ou dependente) e *X* é a variável explicativa (ou independente) e a relação entre elas é linear, pode-se estabelecer a seguinte equação:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

onde β_0 é a intersecção na origem, β_1 é o declive da reta de regressão e ε o termo erro, que representa os fatores que poderão ser determinantes para explicar *Y*, mas não constam da equação. A interpretação dos parâmetros é muito importante, nomeadamente β_1 mede a sensibilidade da variável *Y* a variações na variável *X*, por exemplo se *X* aumentar uma unidade, então *Y* irá variar β_1 unidades. Quanto a β_0, se a variável *X* tiver o valor 0, então a variável *Y* irá assumir o valor de β_0. De salientar que esta interpretação pode ser desprovida de sentido real em alguns casos (por exemplo casos em que não faz sentido que *X* seja igual a 0).

Um dos métodos mais usados para estimar um modelo de regressão linear é o MMQ, pois é o que permite que a soma do quadrado dos desvios seja mínima. Considerando que a soma dos desvios em relação à média é nula e a utilização de módulos provocaria o enviesamento dos resultados, a utilização de quadrados surge como a melhor alternativa. Se forem satisfeitos os pressupostos base deste método de estimação (Teorema de Gauss-Markov), então os estimadores obtidos são considerados *BLUE* (*Best Linear Unbiased Estimators*) e assim, são os estimadores mais eficientes na classe dos estimadores lineares. Estes pressupostos prendem-se, em grande parte, com o comportamento do termo erro, mas não só. De forma resumida podem ser descritos da seguinte forma:

1. Linearidade dos parâmetros;
2. As observações constituem uma amostra aleatória da população;
3. A variável independente *X* não é uma constante;
4. O erro tem valor esperado nulo, dado o valor da variável independente, ou seja, $E(\varepsilon|X) = 0$;
5. Homocedasticidade – o termo erro, ε, tem variância constante, seja qual for o valor de *X*, $Var(\varepsilon|X) = \sigma^2$.

Os primeiros quatro pressupostos garantem o não enviesamento dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$. Já a hipótese de homocedasticidade é a base para a obtenção das equações da variância dos estimadores do MMQ.

De relembrar também a diferença entre erro e resíduo (que tantas vezes são confundidos). Admitindo que há uma função não observável que relaciona a variável independente com a variável dependente, então os desvios das observações da variável dependente desta função são os erros não observáveis (ε). Se for estimado um modelo de regressão para algumas observações (amostra), os desvios das observações da variável dependente face aos valores estimados serão os resíduos $(\hat{\varepsilon})$.

Cada observação gerará um resíduo, que poderá ser positivo se o valor estimado for inferior ao observado, ou negativo no caso oposto. Os resíduos obtidos através do MMQ têm importantes propriedades, nomeadamente: a respetiva soma ser nula (e por isso $\bar{y} = \overline{\hat{y}}$), a covariância amostral entre os resíduos e os regressores ser nula e o ponto $(\bar{X}, \bar{Y})$ estar sempre sobre a linha da regressão.

A análise de variância constitui uma base para avaliação da qualidade de ajustamento do modelo e respetiva significância estatística. Nesta análise de variância, ou ANOVA, é decomposta a soma de quadrados total (SST - *Total Sum of Squares*) em duas parcelas: a soma de quadrados da regressão (SSE - *Explained Sum of Squares*) e a soma de quadrados dos resíduos (SSR - *Residual Sum of Squares*). O poder explicativo do modelo de regressão linear pode ser avaliado através do coeficiente de determinação (R^2), que mede a percentagem da variação da variável dependente que é explicada pelo modelo de regressão.

Importa referir que no caso de se obter um coeficiente de determinação nulo, tal não implica que as variáveis não estejam correlacionadas, simplesmente a relação entre elas pode não ser linear. A significância estatística da equação de regressão (como um todo) pode ser avaliada através da análise de variância, onde é utilizada a estatística F (distribuição F-*Snedcor*).

Finalmente, a inferência individual aos parâmetros do modelo assume elevada importância, pois permite aferir a significância estatística dos mesmos, a importância da variável independente para explicar a variável dependente e ainda inferenciar sobre o comportamento destes parâmetros em contexto mais alargado. Esta é normalmente realizada através de teste de hipótese, recorrendo-se à distribuição t-*Student* para tal. Importa referir que a inferência estatística ao modelo de regressão linear, assume que os resíduos devem ser independentes e identicamente distribuídos (*i.i.d.*), com média nula e variância constante, devendo seguir uma distribuição Normal, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Exemplo 1

Foram selecionados aleatoriamente 269 jogadores de futebol de clubes portugueses da 2.ª divisão. O objetivo é explicar o salário médio mensal auferido por estes jogadores. Assim, foi recolhida a informação acerca do salário médio mensal (*Sal*) dos jogadores e número de anos de experiência enquanto jogadores profissionais (*Exp*) com vista a estimar o seguinte modelo:

$$Sal = \beta_0 + \beta_1 Exp + \varepsilon$$

Através do MMQ foi obtida a seguinte equação:

$$\widehat{Sal}_i = 807{,}93 + 120{,}32 Exp_i$$

$$n = 269$$

$$R^2 = 0{,}167; F_{(1;267)} = 53{,}69\ p - value = 0{,}000$$

Os resultados obtidos permitem-nos concluir que um jogador sem experiência terá um salário médio esperado de, aproximadamente 807,93 euros e que por cada ano de experiência enquanto jogador profissional, o seu salário crescerá cerca de 120 euros. Como podemos ver pela estatística F e respetivo *p-value*, o modelo é globalmente estaticamente significativo, mas explica apenas aproximadamente 16,7% da variação dos salários.

Como podemos melhorar este modelo, no que toca ao nível de explicação da variação dos salários? Uma possibilidade seria incluirmos mais variáveis...

3.2. Motivação para o modelo de regressão linear múltipla, estimação do modelo e sua interpretação

No modelo de regressão linear simples admite-se que o comportamento da variável dependente é influenciado por uma única variável explicativa. O facto de existir uma única variável independente, ou explicativa, é o que confere o modelo a classificação de "simples", característica que é muito desejável quando precisamos modelar um determinado problema, mas que pode revelar insuficiência ou alguma pobreza na explicação dos fenómenos que pretendemos analisar. Com efeito, não são abundantes os exemplos de modelos de regressão linear simples que sejam perfeitamente suficientes (ou até bons) para explicar a variável dependente em estudo. Isto porque, não serão abundantes os fenómenos que precisarão de uma só variável para serem explicados com robustez e qualidade. E é exatamente esta a motivação para a regressão linear múltipla: a necessidade de melhores modelos, a necessidade de melhor explicar a variável dependente, fator que nos impele a procurar mais variáveis explicativas para o modelo.

Inclusivamente, a teoria económica indica que muitos fenómenos podem ser influenciados por diversos fatores, emergindo a necessidade que o modelo integre mais variáveis explicativas. Admite-se, portanto, que a variável dependente Y poderá ser função linear de várias variáveis independentes (ou explicativas) $X_1, X_2, \dots X_k$ e do erro ε:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon.$$

Este modelo tem $k + 1$ parâmetros, onde se incluem todos os coeficientes das k variáveis explicativas, mais o termo constante (β_0). A interpretação dos parâmetros β_j (com $j = 1, \dots, k$) é semelhante à interpretação realizada na análise de regressão simples, ou seja β_0 indica o valor de Y quando todas as variáveis explicativas são nulas (ainda que nem sempre faça sentido que certas variáveis explicativas tomem o valor zero e por isso poder ser desprovido de sentido real) e os demais β_j indicam a sensibilidade de Y quando a respectiva variável explicativa varia uma unidade, mantendo-se todas as outras constantes. O conceito de *ceteris paribus* é aqui muito importante na interpretação isolada dos parâmetros. Os parâmetros β_j são também designados de parâmetros parciais, uma vez que fornecem uma medida da influência de cada uma das variáveis explicativas, assumindo que todas as outras

se mantêm constantes.

O modelo de regressão linear múltipla pode ser apresentado sob a forma de n equações, cada qual correspondendo à observação i de todas as variáveis incluídas no modelo:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + \varepsilon_2$$
$$\ldots$$
$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots + \beta_k X_{kn} + \varepsilon_n.$$

A formulação matricial correspondente é dada por:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

Y e ε são vetores de dimensão $n \times 1$, β é um vetor de dimensão $(k+1) \times 1$, e X é uma matriz de dimensão $n \times (k+1)$. Cada elemento da matriz X tem dois índices: o primeiro refere-se à coluna (variáveis) e o segundo à linha (observação).

Tal como analisado na análise do modelo de regressão simples, o método dos mínimos quadrados tem por objetivo encontrar um vetor de estimadores $\hat{\beta}$ que minimizem a soma dos quadrados dos resíduos (*SSR*). A solução é dada pela seguinte equação:

$$\hat{\beta} = (X'X)^{-1}(X'Y).^{1}$$

Tal como referenciado para os estimadores do MMQ no modelo de regressão linear simples, os estimadores $\hat{\beta}$ são os mais eficientes de entre os estimadores lineares não enviesados para β, ou seja, são os que têm variância mínima e por isso se diz que são *BLUE* (Teorema de Gauss-Markov). Iremos revisitar os pressupostos do MMQ para a regressão linear múltipla no ponto seguinte. Mas retomemos o exemplo iniciado aquando da exploração do modelo de regressão linear simples.

[1] A matriz $X'X$ designa-se por matriz dos produtos cruzados e está garantido que tem inversa pois, de acordo com as hipóteses do modelo de regressão linear múltipla, esta deverá ter característica $k+1$.

Exemplo 2

Partindo do exemplo explorado na regressão linear simples, em que tentamos explicar o salário dos jogadores de futebol da 2.ª Liga em Portugal, foi obtida informação acerca das seguintes variáveis:

Sal – Salário médio mensal do jogador

Exp – Número de anos de experiência enquanto jogador profissional

Jogos – Número de jogos em que o jogador participou

Minutos – Número total de minutos que o jogador jogou

Golos – Número de golos marcados pelo jogador

Assist – Número de assistências a golos promovidas pelo jogador

Idade – Idade do jogador, em anos

O modelo a estimar é o seguinte:

$$Sal = \beta_0 + \beta_1 Exp + \beta_2 Jogos + \beta_3 Minutos + \beta_4 Golos + \beta_5 Assist + \beta_6 Idade + \varepsilon$$

Através do MMQ, foram obtidos os seguintes resultados:

$$\widehat{Sal}_i = -214{,}78 + 76{,}70 Exp_i - 2{,}75 Jogos_i - 0{,}042 Minutos_i + 89{,}91 Golos_i + 81{,}78 Assist_i + 7{,}68_6 Idade_i$$

$$n = 269$$

Numa primeira análise (e ainda sem avaliar a significância estatística de cada parâmetro), vemos que as variáveis que influenciam positivamente o salário serão a experiência, o número de golos e de assistência e a idade do jogador. O valor de $\hat{\beta}_0 = -214{,}78$ indicaria que o salário de um jogador que tivesse o valor zero em todas as variáveis seria de -214,78 euros, ou seja, um jogador recém-nascido, sem experiência, com zero jogos e zero minutos jogados, que nunca tenha marcado golos nem feito assistências, paga 214,78 euros para trabalhar. Ora esta interpretação é desprovida de sentido real, uma vez que existem variáveis (nomeadamente a idade) que não faz sentido serem nulas neste modelo.

A interpretação dos restantes estimadores deve respeitar a regra de *ceteris paribus*, por exemplo, no que se refere ao coeficiente da variável *Exp*, podemos referir que, mantendo tudo o resto constante, por cada ano a mais de experiência, é expectável que o salário mensal médio do jogador aumente 76,70 euros.

3.3. Pressupostos para a regressão linear múltipla

É importante conhecer as propriedades estatísticas dos estimadores do MMQ. Neste ponto, é derivado o valor esperado e variância destes estimadores, tendo por base os pressupostos do Teorema de Gauss-Markov. São aqui apresentados os pressupostos para que os estimadores sejam não enviesados, pressupostos estes que não são mais que a extensão dos pressupostos dos estimadores do MMQ para o modelo de regressão linear simples. A hipótese de eficiência é também explorada, tendo por base o comportamento da variância do erro.

O primeiro pressuposto define, simplesmente, o modelo de regressão linear múltipla:

Pressuposto 1 – Linearidade dos parâmetros

Este pressuposto admite que o modelo de regressão linear múltipla pode ser definido da seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon.$$

Onde $\beta_0, \beta_1, \ldots, \beta_k$ são os parâmetros não conhecidos, mas constantes e ε é o erro aleatório, ou os desvios. O fator chave nesta equação é a linearidade nos parâmetros, podendo as variáveis dependente e independentes constituir funções não lineares das variáveis de interesse (por exemplo logaritmos, polinómios, etc.).

Pressuposto 2 – Amostra aleatória

Para a estimação do modelo apresentado no Pressuposto 1, é utilizada uma amostra aleatória com *n* observações. Se quisermos particularizar o modelo global para uma específica observação *i*, teremos:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i.$$

Pressuposto 3 – Ausência de colinearidade

O terceiro pressuposto impõe que nenhuma variável independente seja uma constante e que não exista colinearidade perfeita entre as variáveis independentes.

Relativamente ao modelo de regressão linear simples, esta é uma hipótese nova no modelo de regressão linear múltipla.

A existência de colinearidade perfeita impediria a estimação do modelo através do MMQ, uma vez que o determinante da matriz $X'X$ seria nulo, impedindo de se encontrar $(X'X)^{-1}$. Não existindo colinearidade perfeita, o modelo de regressão linear é estimável, mas há que ter em conta a possível relação linear (ainda que não perfeita) entre as variáveis explicativas, ou seja, ter em conta a existência de multicolinearidade que poderá afetar a eficiência dos estimadores.

A colinearidade perfeita pode ocorrer por diversas causas, nomeadamente a inclusão da mesma variável com diferentes unidades de medida, ou a inclusão de uma variável que é resultado da função linear entre outras variáveis (veja-se o exemplo em que $X_3 = X_1 + X_2$, onde X_1 reflete o valor gasto na campanha publicitária para livros em formato tradicional, X_2 o valor da campanha publicitária para *e-books* e X_3 o total gasto em campanhas publicitárias para livros por parte da empresa XPTO).

Pressuposto 4 – O valor esperado do erro, dado o comportamento das variáveis independentes, é nulo

Este pressuposto indica que o valor esperado do erro ou dos desvios deve ser nulo, não estando correlacionado com as variáveis independentes:

$$E(\varepsilon|X_1, X_2, \dots, X_k) = 0$$

A violação deste pressuposto poderá estar relacionada com problemas de especificação do modelo, omissão de variáveis importantes para o modelo e endogeneidade de variáveis independentes. A análise destes problemas não constitui um tópico da unidade curricular de Análise de Dados para Negócios I, contudo considero importante alertar os estudantes para tal e referenciar bibliografia que os poderá ajudar, nomeadamente Wooldridge (2019), capítulos 9 e 16.

Os pressupostos 1 a 4 permitem conhecer o valor esperados dos estimadores $\hat{\beta}_j$, mas é necessária a avaliação da dispersão. Tal como já analisado no caso do modelo de regressão

linear simples, a homocedasticidade é um pressuposto fundamental para garantir a eficiência destes estimadores e no caso do modelo de regressão linear múltipla, a sua importância é igualmente inquestionável.

Pressuposto 5 – Homocedasticidade

Este pressuposto assume que a variância do termo erro é constante, independentemente do comportamento das variáveis independentes:

$$Var(\varepsilon|X_1, X_2, \dots, X_k) = \sigma^2$$

Quando este pressuposto não é verificado, estamos perante um modelo que exibe heterocedasticidade. Um exemplo é a análise dos resultados líquidos e capitais próprios de empresas de diferente dimensão, onde é possível a ocorrência de um aumento em termos de dispersão dos resultados à medida que aumenta a dimensão da empresa.

Admitindo que existe heterocedasticidade, mas que o valor esperado do erro é zero e que este não está correlacionado com as variáveis independentes, as consequências da violação desta hipótese são de diversa ordem. Os estimadores do MMQ continuam a ser não enviesados, mas deixam de ser eficientes, pois passa a ser possível encontrar estimadores lineares centrados com menor variância em comparação com os estimadores do MMQ. Assim sendo, os estimadores deixam de ser *BLUE* e a inferência estatística para os mesmos perde robustez, uma vez que maiores valores de variância tornarão menos precisos os estimadores, o que se traduz em intervalos de confiança de maior amplitude e testes de hipóteses menos precisos.

3.4. Avaliação do ajustamento e inferência no modelo de regressão

3.4.1. O coeficiente de determinação e o teste à significância global do modelo

Tal como no modelo de regressão linear simples, pode-se decompor a variabilidade de Y em duas componentes de variabilidade: a variabilidade explicada pela regressão e a variabilidade residual.

Novamente, como referido aquando do modelo de regressão linear simples, o

coeficiente de determinação pode variar entre 0 e 1 e quanto maior é o seu valor, maior é o poder explicativo do modelo de regressão em estudo. O problema do coeficiente de determinação R^2, como medida de qualidade do ajustamento, reside no facto de este se referir apenas à variação explicada de Y e não levar em consideração os graus de liberdade que lhe estão associados. Uma alternativa é considerar as variâncias em vez das variações, resultando no coeficiente de determinação corrigido ou coeficiente de determinação ajustado $(\bar{R}^2)$, eliminando-se assim a dependência da qualidade do ajustamento do número de variáveis explicativas que fazem parte do modelo (de relembrar que a variância é igual à variação dividida pelos graus de liberdade). Neste caso:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}.$$

Sempre que novas variáveis são introduzidas no modelo, há que verificar:

1) A soma dos quadrados do dos resíduos diminui ou permanece inalterada e R^2 eleva-se ou permanece constante;
2) Se o valor de $\bar{R}^2$ baixar, significa que o ganho em R^2 é insuficiente para compensar a perda de graus de liberdade que lhe está associada, resultando numa diminuição da qualidade do ajustamento.

Tal como foi visto para a regressão linear simples, a tabela ANOVA para a regressão linear múltipla é definida da seguinte forma:

Tabela 1. ANOVA aplicada ao modelo de regressão linear múltipla.

Fonte da variação	*Soma de quadrados*	*g.l.*	*Quadrado médio*
Regressão (*explained*)	$SSE=\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	$(k+1)-1$	$MSE = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{k}$
Resíduos	$SSR=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$	$(n-k-1)$	$MSR = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-k-1}$
Total	$SST=\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$(n-1)$	

Nota: g.l. – graus de liberdade e *k* é o número de variáveis independentes.

Para podermos avaliar a significância global do modelo, as hipóteses em estudo são as seguintes:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \exists\, \beta_j \neq 0$$

A realização deste teste inclui a utilização da estatística F, que avalia a significância estatística do modelo como um todo, utilizando a base da construção do coeficiente de determinação R^2. A estatística F, com k e $n-k-1$ graus de liberdade, permite testar se, simultaneamente, nenhuma das variáveis independentes contribui para explicar a variação de Y em relação à sua média.

A estatística do teste é, assim, definida por

$$F_{k,n-k-1} = \frac{\frac{SSE}{k}}{\frac{SSR}{n-k-1}} = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$$

Exemplo 3

Retomemos o exemplo em estudo, cujos resultados obtidos foram:

$$\widehat{Sal}_i = -214{,}78 + 76{,}70Exp_i - 2{,}75Jogos_i - 0{,}042Minutos_i + 89{,}91Golos_i + 81{,}78Assist_i + 7{,}68_6 Idade_i$$

$$n = 269$$

$$R^2 = 0{,}554;\ \bar{R}^2 = 0{,}544;\ F_{(6;262)} = 54{,}263\ p-value = 0{,}000$$

Verificamos que o modelo tem a capacidade de explicar 55,4% da variação dos salários. Para avaliarmos as hipóteses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_6 = 0$$
$$H_1 : \exists\, \beta_j \neq 0$$

Utilizamos a estatística *F*, cujo valor é 54,26 e respetivo *p-value*=0,000, concluindo-se que o modelo é globalmente significativo.

Mas serão todas as variáveis estatisticamente significativas?

3.4.2. Inferência sobre um só parâmetro

Para se poder proceder à análise inferencial (intervalos de confiança e testes de

hipóteses) dos parâmetros na análise de regressão múltipla para parâmetros individuais, o processo é muito semelhante ao apresentado para a regressão linear simples. Sabendo que a variância dos estimadores é dada por:

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1},$$

Então, espera-se que estes sigam o seguinte comportamento:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}),$$

sendo $S^2_{\hat{\beta}}$ o estimador para a variância de $\hat{\beta}$, podemos obter a estatística seguinte:

$$T = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{n-k-1}$$

Que segue uma distribuição *t-Student* com $n-k-1$ graus de liberdade. Esta estatística pode ser usada para testar, por exemplo, a significância de um determinado parâmetro ($H_0: \beta_j = 0$)[2] ou um determinado valor para o parâmetro em causa. O teste seria formulado da seguinte forma:

$$H_0: \beta_j = \beta_j^0,$$

e as hipóteses alternativas

$$H_1: \beta_j \neq \beta_j^0 \Rightarrow P\left(T \leq -t_{n-k-1,\alpha/2} \text{ ou } T \geq t_{n-k-1,\alpha/2}\right) = \alpha$$
$$H_1: \beta_j < \beta_j^0 \Rightarrow P\left(T \leq -t_{n-k-1,\alpha}\right) = \alpha$$
$$H_1: \beta_j > \beta_j^0 \Rightarrow P\left(T \geq t_{n-k-1,\alpha}\right) = \alpha.$$

Para além de testes de hipótese, podemos construir intervalos de confiança:

$$P\left(\hat{\beta}_j - t_{\alpha/2,(n-k-1)}S_{\hat{\beta}_i} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,(n-k-1)}S_{\hat{\beta}_i}\right) = 1-\alpha,$$

$$\left]\hat{\beta}_j - t_{\alpha/2,(n-k-1)}S_{\hat{\beta}_i};\ \ \hat{\beta}_j + t_{\alpha/2,(n-k-1)}S_{\hat{\beta}_i}\right[.$$

[2]De salientar que se se testar a significância de um parâmetro, o valor da estatística $T = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$ depende do valor da estimativa $\hat{\beta}_j$ e do seu desvio-padrão. Se o estimador for muito preciso, ou seja, se $S_{\hat{\beta}_j}$ for muito pequeno, é natural que se rejeite a hipótese nula $H_0: \beta_j = 0$, mesmo que $\hat{\beta}_j$ tenha um valor próximo de zero. É natural que em pequenas amostras, este desvio-padrão tenda a ter valores mais elevados. De salientar também a possível influência da multicolinearidade sobre este estimador, que poderá elevar o seu valor.

Exemplo 4

Voltando ao exemplo explorado neste documento, comecemos por avaliar a significância estatística de cada estimador, ou seja:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

A Tabela seguinte (retirada diretamente do output do Excel) apresenta os resultados para os estimadores, respetivo desvio-padrão, estatística *t-Student* e *p-valu*e para as hipóteses formuladas. São ainda apresentados os limites inferior e superior para os intervalos de confiança para cada estimador (assumindo um nível de confiança de 95%).

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	-214,778	859,731	-0,250	0,803	-1907,640	1478,084
Exp	**76,703**	**37,464**	**2,047**	**0,042**	**2,935**	**150,471**
Jogos	-2,754	4,151	-0,663	0,508	-10,928	5,420
Minutos	-0,042	0,144	-0,295	0,768	-0,325	0,240
Golos	**89,912**	**14,790**	**6,079**	**0,000**	**60,790**	**119,035**
Assist	**81,775**	**18,231**	**4,485**	**0,000**	**45,876**	**117,673**
Idade	7,680	37,099	0,207	0,836	-65,371	80,730

Como podemos verificar (e assumindo um nível de significância de 5%), apenas três parâmetros são estatisticamente diferentes de zero, o que indica que são estatisticamente significativas enquanto determinantes do salário dos jogadores de futebol, as variáveis: experiência, golos e assistências. Se observarmos os respetivos intervalos de confiança, verificamos que estes são os que não incluem o valor zero.

Mas os testes aos parâmetros individuais não se esgotam na análise da respetiva significância estatística. Podemos realizar outros testes. Por exemplo, imagine que um famoso comentador desportivo afirma que na 2.ª Liga de Portugal o impacto de cada golo marcado pelo jogador no seu salário é de pelo menos 100 euros. Como podemos testar esta afirmação?

As hipóteses em estudo seriam:

$$H_0: \beta_4 \leq 100$$

$$H_1: \beta_4 > 100$$

Neste caso, a estatística *t* seria dada por:

$$T = \frac{89{,}912 - 100}{14{,}79} \sim T_{268;0{,}05}$$

$$T = -0{,}682$$

Temos então um teste unilateral à direita. Se assumirmos um nível de significância de 5%, verificamos na tabela da distribuição *t*-Student que o valor crítico $t_{268;0{,}05} = 1{,}645$, sendo definidas as regiões de não rejeição e de rejeição da seguinte forma:

$$RNR =]-\infty;\ 1{,}645[$$

$$RR = [1{,}645;\ +\infty[$$

Verificamos que o valor da estatística *T* pertence à região de não rejeição, ou seja, não rejeitamos a hipótese nula de igualdade de que o impacto de mais um golo no salário é inferior, ou quanto muito igual, a 100 euros. Por outras palavras, o comentador não tem razão (assumindo um nível de significância de 5%)

3.4.3. Teste sobre um subconjunto de parâmetros

Para testar a aderência global do modelo é utilizado o teste F, tal como já foi apresentado. Contudo pode haver interesse em testar apenas uma parte do modelo, saber se um conjunto de variáveis têm ou não significância. Por exemplo, imaginemos que o modelo inicial tem k variáveis explicativas, mas pretende-se analisar a significância de conjunta k_1 variáveis, sendo $k_1 < k$. No fundo, pretende-se perceber se um modelo mais restrito, poderá ser mais adequado para explicar a variável dependente.

A hipótese nula a testar é:

$$H_0:\ \beta_1 = \beta_2 = \ldots = \beta_{k_1} = 0.$$

Se a hipótese nula não for rejeitada, então o modelo pode ser escrito por

$$Y_i = \beta_0 + \beta_{k_1+1} X_{k_1+1,i} + \beta_{k_1+2} X_{k_1+2,i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$

incluindo apenas as variáveis $k - k_1$.

Para a realização do teste há que estimar a soma dos quadrados dos resíduos (SSR^*) referente ao novo modelo (chamado de modelo restrito), obter a soma dos quadrados dos resíduos do modelo global (SSR) e calcular a seguinte estatística:

$$F = \frac{\frac{SSR^* - SSR}{k_1}}{\frac{SSR}{n-k-1}} \sim F_{k_1, n-k-1}.$$

A intuição por detrás deste teste é que se a hipótese nula é verdadeira, então SSR^* e SSR não divergem muito. Se a diferença entre estes for muito grande ($SSR < SSR^*$), quer dizer que a inclusão de todas as variáveis no modelo leva a que a soma dos quadrados dos

resíduos diminua o bastante para rejeitar a exclusão de certas variáveis.

Exemplo 5

Tendo por base os resultados obtidos no modelo global para explicar os salários dos jogadores, e tendo em consideração que vários parâmetros não se mostraram significativos, poderíamos tentar avaliar se a sua eliminação seria benéfica para o modelo. Assim teríamos as hipóteses:

$$H_0: \quad \beta_2 = \beta_3 = \beta_6 = 0$$
$$H_1: \exists\, \beta_j \neq 0$$

Haveria então necessidade de estimar um novo modelo, o modelo restrito:

$$Sal = \beta_0 + \beta_1 Exp + \beta_4 Golos + \beta_5 Assist + \varepsilon$$

Através do MMQ, foram obtidos os seguintes resultados:

$$\widehat{Sal}_i = -179{,}35 + 82{,}73 Exp_i + 80{,}81 Golos_i + 71{,}46 Assist_i$$
$$n = 269$$
$$R^2 = 0{,}550;\ \bar{R}^2 = 0{,}545; F = 108{,}95\ p - value = 0{,}000$$

Como podemos verificar, os sinais dos estimadores não se alteraram, o coeficiente de determinação baixou ligeiramente, mas o $\bar{R}^2$ aumentou ligeiramente, o que, a priori, já nos indica que as três variáveis eliminadas não deverão ser muito importantes no modelo. Vejamos então a estatística F:

$$F = \frac{\frac{120464406{,}638 - 119446231{,}845}{3}}{\frac{119446231{,}845}{269 - 6 - 1}} \sim F_{3,262}.$$
$$F = 0{,}744$$

Assumindo um nível de significância de 5%, as regiões de não rejeição e de rejeição são as seguintes:

$$RNR = [0; 2{,}60[$$
$$RR = [2{,}60;\ +\infty[$$

Concluímos que a hipótese nula não deverá ser rejeitada, o que indica que o modelo restrito é melhor que o modelo global.

3.5. Modelo de regressão linear múltipla com informação qualitativa

Até agora, não demos especial importância à existência de variáveis qualitativas enquanto variáveis independentes. Estas variáveis podem ter um contributo muito importante e deverão ser construídas e analisadas com alguns cuidados. Numa primeira fase, há que distinguir entre os diferentes tipos de variáveis qualitativas que se podem utilizar.

3.5.1. Variáveis binárias, dicotómicas ou dummy

Estas variáveis tomam o valor 0 ou 1 consoante um de dois resultados ocorre. As variáveis *dummy* são, das variáveis discretas, aquelas que são mais utilizadas em trabalhos aplicados. Se se pretender estudar a diferenciação salarial entre homens e mulheres de uma dada categoria, numa certa empresa, a variável *dummy* a utilizar seria, por exemplo

$$D_i = \begin{cases} 0 & \textit{se é sexo masculino} \\ 1 & \textit{se é sexo feminino} \end{cases}.$$

Retomando o exemplo em análise, poderia ser interessante incluir variáveis qualitativas, no formato de variáveis binárias (variáveis *dummy*). Por exemplo, poderia ser interessante incluir uma variável respeitante à existência no passado de lesões graves. Esta variável poderia ser um determinante importante, a par da experiência do jogador. Admitindo que a existência de informação sobre lesões graves era apenas sim/não, seria modelada de modo binário: 1 se já teve ou tem lesões graves, 0 se não teve ou não tem.

$$Lesoes_i = \begin{cases} 1 \ \textit{se} \text{ jogador i já teve ou tem lesões graves} \\ 0 \ \textit{se} \text{ jogador i não teve ou não tem lesões graves} \end{cases}$$

3.5.1.1. *Dummy*: alteração na intersecção na origem

Podem existir situações em que a variável *dummy* apenas influencia a intersecção na origem. Esta ação é designada, também, de inclusão de variáveis *dummy* na forma aditiva. O modelo pode ser descrito por:

$$Y_i = \beta_0 + \delta D_i + \beta_1 X_i + \varepsilon_i.$$

Neste caso, a variável *dummy* irá influenciar a parte constante da regressão e o resultado desta parte será

$$\begin{cases} \beta_0 \text{ se } D_i = 0 \\ \beta_0 + \delta \text{ se } D_i = 1 \end{cases}$$

Nesta situação, o coeficiente δ indica o deslocamento na intersecção na origem, mantendo-se constante o declive.

Exemplo 6

Vamos então incluir a variável binária Lesoes no nosso modelo. Teremos então o seguinte modelo a estimar:

$$Sal = \beta_0 + \beta_1 Exp + \beta_2 Golos + \beta_3 Assist + \delta_1 Lesoes + \varepsilon$$

Para uma mais fácil interpretação, podemos apresentar da seguinte forma:

$$\begin{cases} \widehat{Sal}_i = -177{,}84 + 83{,}14 Exp_i + 80{,}96 Golos_i + 78{,}59 Assist_i; se\ Lesoes = 0 \\ \widehat{Sal}_i = -177{,}84 + 83{,}14 Exp_i + 80{,}96 Golos_i + 78{,}59 Assist_i - 16{,}88;\ se\ Lesoes = 1 \end{cases}$$

Ou seja, para jogadores com as mesmas características, estima-se que um jogador que tem ou teve lesões graves, tenha uma redução no seu salário médio de 16,88 euros.

O coeficiente δ_1 pode ser alvo de teste e avaliada a sua significância. As hipóteses a formular neste caso são

$$H_0 : \delta_1 = 0$$

$$H_1 : \delta_1 \neq 0.$$

Este é um teste a um parâmetro individual, não existindo diferença alguma nos procedimentos pelo facto de a variável que lhe está associada ser qualitativa.

3.5.1.2. *Dummy*: alteração do declive

A inclusão de variáveis *dummy* na forma multiplicativa pode promover alterações no declive da regressão. O modelo para este tipo de situações pode definir-se do seguinte modo:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma D_i X_i + \varepsilon_i,$$

ou seja

$$Y_i = \beta_0 + (\beta_1 + \gamma D_i) X_i + \varepsilon_i.$$

Exemplo 7

Consideremos o Exemplo 6. Admita que após uma análise mais cuidada do modelo e respetivas variáveis, se concluiu que possivelmente a existência de lesões influenciaria o salário médio dos jogadores, mas de forma dependente da experiência dos mesmos.

O modelo a estimar seria:

$$Sal = \beta_0 + \beta_1 Exp + \beta_2 Golos + \beta_3 Assist + \delta_1 Lesoes.Exp + \varepsilon$$

Ou seja

$$Sal = \beta_0 + (\beta_1 + \delta_1 Lesoes) Exp + \beta_2 Golos + \beta_3 Assist + \varepsilon$$

Foram obtidos os seguintes resultados:

$$\widehat{Sal}_i = -191{,}57 + (88{,}71 - 12{,}42 Lesoes_i) Exp_i + 81{,}75 Golos_i + 78{,}01 Assist_i$$

$$n = 269$$

$$R^2 = 0{,}552;\ \bar{R}^2 = 0{,}545; F = 81{,}22\ p - value = 0{,}000$$

Se avaliarmos a derivada do salário estimado ($\widehat{Sal}$) face à experiência (*Exp*), obtemos:

$$\frac{\partial \widehat{Sal}}{\partial Exp} = 88{,}81 - 12{,}42 Lesoes$$

Ou seja, há uma alteração do declive ou do impacto da variável *Exp* no salário médio dos jogadores, que aquando da existência de lesões diminui em 12,42 euros, ou seja, passa a ser 76,39 euros.

3.5.2. Variáveis policotómicas

Estas variáveis podem assumir, numa numeração discreta, mais que dois valores possíveis, podendo ser de dois diferentes tipos:

Variáveis não ordinais, nas quais não há uma ordenação natural das alternativas. Admitamos, por exemplo, que nos interessa incluir a posição usual do jogador em campo, para perceber se esta variável é um determinante do salário. Assumindo que existem 3 posições: "Defesa", "Guarda-redes" e "Avançado", poderíamos ter a seguinte variável policotómica:

$$\mathrm{Pos_i} = \begin{cases} 1 \text{ se o jogador i for Defesa} \\ 2 \text{ se o jogador i for Guarda} - \text{redes} \\ 3 \text{ se o jogador i forAvançado} \end{cases}$$

Neste caso, o valor atribuído a cada possibilidade não tem qualquer ordem incluída, trata-se apenas de uma diferenciação nominal, ou seja, o valor um não indica um menor atributo face ao dois e ao três. A utilização de uma variável policotómica com os valores um, dois e três para representar a posição do jogador no campo num modelo de regressão, poderia não ser correta, dado que o valores numéricos são desprovidos de sentido real, são apenas uma codificação nominal. A Tabela 2 apresenta a construção das novas variáveis *dummy*:

Tabela 2. Modelação de uma variável policotómica nominal em variáveis *dummy*

Posição do jogador	Pos_1	Pos_2	Pos_3
Defesa	1	0	0
Guarda-redes	0	1	0
Avançado	0	0	1

Onde Pos_1, Pos_2 e Pos_3 são as variáveis *dummy* para o descrever a posição do jogador em campo. Se se quiser modelar o salário (Sal) através da posição do jogador, então teríamos:

$$Sal = \beta_0 + \delta_1 Pos_1 + \delta_2 Pos_2 + \varepsilon$$

É de referenciar que o número de variáveis *dummy* a criar para representar os diversos atributos (p) assumidos por uma variável de tipo qualitativo será igual a $p - 1$. Por exemplo, neste caso a variável podia assumir três diferentes valores e por isso basta usar duas variáveis *dummy*. Se se conhecerem $p - 1$ variáveis, sabe-se automaticamente o valor da p -ésima variável. Caso fossem incluídas p variáveis *dummy* no modelo, incorreríamos em problemas de colinearidade e uma variável poderia ser encontrada enquanto combinação linear das restantes, pois $\sum_{i=1}^{3} Pos_i = 1$, e assim $D_3 = 1 - D_1 - D_2$, condição impeditiva para a obtenção de uma solução única na resolução do sistema que permite o cálculo através do MMQ.[3]

Também podemos ter variáveis policotómicas em que o valor atribuído tem, em si, uma ordem. São as *Variáveis ordinais*, onde existe uma ordem natural. Por exemplo se quiséssemos avaliar o nível de motivação dos jogadores, poderíamos ter uma variável definida por:

[3]Esta situação é também habitualmente designada por *"armadilha das variáveis dummy"*.

$$Mot_i = \begin{cases} 1 \textit{ se o jogador i não está motivado} \\ 2 \textit{ se o jogador i está razoavelmente motivado} \\ 3 \textit{ se o jogador i está muito motivado} \end{cases}$$

Este tipo de variável pode ser incluído no modelo de regressão com os seus valores originais, uma vez que os valores numéricos representam uma ordem.

$$Sal_i = \beta_0 + \beta_1 Mot_i + \varepsilon_i$$

Ainda assim, há que ter alguns aspetos em consideração, pois não se trata de uma variável quantitativa. Quando obtemos o valor de β_1, temos o valor que o salário (variável dependente) irá variar aquando da variação unitária na Motivação. Mas será que passar de um para dois é similar a passar de dois para três? Será que esta escala consegue captar a complexidade do conceito adjacente? Se tivermos dúvidas, talvez seja preferível optar pela utilização de variáveis *dummy* em substituição da variável original:

$$Sal = \beta_0 + \delta_1 Mot_1 + \delta_2 Mot_2 + \varepsilon$$

4. CONSIDERAÇÕES FINAIS

O objetivo global desta Lição é a apresentação e compreensão do modelo de regressão linear múltipla enquanto ferramenta que pode apoiar o processo de tomada de decisão em gestão. Deste modo, a interpretação dos parâmetros reveste-se de elevada importância. Naturalmente que a compreensão do método para a obtenção dos estimadores é essencial, assim como dos seus pressupostos. No que toca aos pressupostos, é de referir que os mesmos são analisados com maior detalhe aquando da apresentação do modelo de regressão linear simples, sendo dado maior realce nesta Lição àquele que é incluído de novo: a ausência de colinearidade. Nesta Lição não se procederá à verificação dos mesmos, mas sim à sua explanação e realce da importância de os observar, de modo a obter resultados válidos e robustos. A respetiva verificação é realizada em aula posterior.

Como facilmente se compreenderá, quaisquer decisões ou conclusões retiradas com os resultados obtidos através de um modelo de regressão linear carecem de verificação da respetiva significância estatística. Por este motivo são explorados testes de hipótese ao modelo global, a um subconjunto dos parâmetros e a parâmetros individuais.

Finalmente, é explorada de forma introdutória a inclusão de variáveis dicotómicas enquanto variáveis independentes, tanto na forma aditiva da interseção na origem, como na forma multiplicativa com variáveis quantitativas.

Acima de tudo pretende-se que os estudantes compreendam a importância desta ferramenta e a explorem de forma correta em trabalhos de investigação onde é necessário conhecer os determinantes de uma determinada variável dependente quantitativa e seccional.

5. REFERÊNCIAS BIBLIOGRÁFICAS

- Wooldridge, J. (2019). *Introductory Econometrics – A Modern Approach*, 7th edition, South-Western College Publishing, Thomson Learning. Florence.