

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



## **Fatores Determinantes na Satisfação Laboral: Um Estudo Quantitativo**

David Henrique Rique Alcazar

**Mestrado em Matemática Aplicada à Economia e Gestão**

Trabalho de Projeto orientado por:

Fernando José Araújo Correia da Ponte Sequeira





# Agradecimentos

Gostaria primeiramente de agradecer ao professor Fernando Sequeira, orientador deste projeto, por todo o apoio dado tanto a nível técnico como pessoal, dando a motivação extra quando estava mais desanimado com este projeto.

À Susana Justo, gestora da consultora Qmetrics, não só pela disponibilização dos dados, mas também por ter assumido uma posição prestativa.

A todos os meus amigos que estiverem comigo neste ciclo de vida, com um agradecimento especial ao André Freitas que esteve comigo tanto na licenciatura como no mestrado. As nossas sessões de estudo, os trabalhos em conjunto e o companheirismo de sala de aula serão recordados com uma grande nostalgia certamente.

Por último agradecer à minha família por todo o apoio dado desde sempre.

## Resumo

A satisfação laboral desempenha um papel crucial no contexto organizacional. Inúmeros estudos têm vindo a mostrar os impactos positivos que a satisfação laboral tem no desempenho, na lealdade, na vontade de permanência, entre outros.

Sendo a satisfação laboral uma peça fulcral no contexto organizacional, torna-se essencial compreender os fatores que levam a um indivíduo sentir-se satisfeito. É neste contexto que surge este estudo.

Os dados utilizados são provenientes de um questionário realizado aos colaboradores de uma empresa inserida no setor indústria/ retalho, no âmbito de um projeto do Observatório Nacional dos Recursos Humanos (ONRH) em parceria com a Qmetrics.

Fez-se uma análise exploratória dos dados e criou-se um modelo de regressão linear múltipla para explicar a satisfação global com a empresa. As variáveis de estudo foram Infraestrutura Laboral, Envolvimento organizacional, Orientação para o cliente, Relacionamentos interpessoais, Autonomia, Reconhecimento, Desenvolvimento profissional, Remuneração/ benefícios e outras variáveis de caracterização.

**Palavras-Chave:** Satisfação laboral; Regressão linear múltipla; Teste de Kruskal-Wallis; Teste de Dunn.

## **Abstract**

Job satisfaction plays a crucial role in the organizational context. Numerous studies have shown the positive impacts that job satisfaction has on performance, loyalty, willingness to stay, among others.

As job satisfaction is a key component in the organizational context, it is essential to understand the factors that lead an individual to feel satisfied. It is in this context that this study emerges.

The data used are from a questionnaire conducted with employees of a company in the industry/retail sector, as part of a project by the National Human Resources Observatory (ONRH) in partnership with Qmetrics.

An exploratory data analysis was performed, and a multiple linear regression model was created to explain overall satisfaction with the company. The study variables were Labor Infrastructure, Organizational Involvement, Customer Orientation, Interpersonal Relationships, Autonomy, Recognition, Professional Development, Compensation/Benefits, and other characterization variables.

**Keywords:** Job satisfaction; Multiple linear regression; Kruskal-Wallis test; Dunn's test



# Índice

AGRADECIMENTOS .....	II
RESUMO.....	III
ABSTRACT .....	IV
ÍNDICE DE FIGURAS .....	VII
ÍNDICE DE TABELAS .....	VII
LISTA DE ABREVIATURAS.....	VII
<b>1. INTRODUÇÃO .....</b>	<b>1</b>
<b>2. OBJETIVOS.....</b>	<b>1</b>
<b>3. RESUMO TEÓRICO .....</b>	<b>2</b>
3.1 DA REGRESSÃO LINEAR MÚLTIPLA AO MODELO LINEAR GENERALIZADO .....	2
3.2 REGRESSÃO LINEAR SIMPLES .....	2
3.3 REGRESSÃO LINEAR MÚLTIPLA.....	7
3.3.1 <i>Estimadores dos parâmetros e as suas propriedades</i> .....	7
3.3.2 <i>Inferência</i> .....	8
3.3.3 <i>Medidas de ajustamento e teste sobre a significância global do modelo</i> .....	9
3.3.4 <i>Escolha do modelo</i> .....	10
3.3.5 <i>Exigências e recomendações do modelo de regressão linear múltipla</i> .....	11
3.3.6 <i>Modelo linear generalizado</i> .....	13
3.4 TESTE DE KRUSKAL-WALLIS .....	14
3.4.1 <i>Teste de post hoc de Dunn</i> .....	15
3.5 COEFICIENTE ALFA DE CRONBACH .....	15
<b>4. METODOLOGIA.....</b>	<b>16</b>
4.1 QUESTIONÁRIO .....	16
4.2 PROCEDIMENTO.....	17
<b>5. RESULTADOS .....</b>	<b>18</b>
5.1 ANÁLISE EXPLORATÓRIA DOS DADOS .....	18
5.1.1 <i>Variáveis de caracterização</i> .....	18
5.1.2 <i>Indicadores</i> .....	19
5.2 ANÁLISE COMPARATIVA.....	20
5.2.1 <i>Teste Kruskal-Wallis – Post Hoc de Dunn</i> .....	21
5.3 APLICAÇÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA .....	24
5.3.1 <i>Avaliação do modelo</i> .....	27
<b>CONCLUSÃO .....</b>	<b>29</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS.....</b>	<b>30</b>

## Índice de figuras

Figura 1 - Valor médio/ desvio padrão por indicador .....	20
Figura 2 - Legenda indicadores .....	22
Figura 3 - Diferenças por Género .....	22
Figura 4 – Diferenças por Habilitações literárias .....	22
Figura 5 - Diferenças por Antiguidade na organização .....	23
Figura 6 - Diferenças por Classe etária .....	23
Figura 7 - Diferenças por Posição hierárquica .....	23
Figura 8 – Representação esquemática do modelo de regressão linear múltipla .....	24
Figura 9 - Histograma resíduos do modelo final .....	27

## Índice de tabelas

Tabela 1 - Tabela Anova .....	6
Tabela 2 - Caracterização dos colaboradores .....	18
Tabela 3 - Coeficientes alfa de Cronbach .....	19
Tabela 4 - Valor médio por estrato .....	21
Tabela 5 - Pontos influentes do primeiro modelo .....	25
Tabela 6 - Teste Kolmogorov-Smirnov aos resíduos do primeiro modelo .....	25
Tabela 7 – Estatísticas de mudança dos modelos criados .....	26
Tabela 8 - Tabela Anova do modelo final .....	26
Tabela 9 - Resumo do modelo final .....	26
Tabela 10 - Teste Kolmogorov-Smirnov aos resíduos do modelo final .....	28
Tabela 11 - Teste Durbin-Watson aos resíduos .....	28

## Lista de Abreviaturas

AIC	<i>Akaike Information Criterion</i> (Critério de Informação de Akaike)
Aut	Autonomia
BLUE	<i>Best Linear Unbiased Estimator</i>
DPro	Desenvolvimento Profissional
EOrg	Envolvimento Organizacional
ILab	Infraestrutura Laboral
LIFCR	(Limite Inferior de Fréchet-Cramer-Rao)
MRLM	Modelo de Regressão Linear Múltipla
MRLS	Modelo de Regressão Linear Simples
OClie	Orientação para o Cliente
ONRH	Observatório Nacional dos Recursos Humanos
RBen	Remuneração/ Benefícios
Rec	Reconhecimentos
RInter	Relacionamentos Interpessoais
SGblo	Satisfação Global
VIF	<i>Variance Inflation Factor</i> (Fator de Inflação da Variância)

# 1. Introdução

Os estudos científicos sobre a satisfação laboral começaram a intensificar-se na década de 1930, maioritariamente influenciado pela Grande Depressão, um período negro da história económica que começou nos Estados Unidos e se estendeu pelo resto do mundo (ScienceDirect. Job Satisfaction). O estudo da satisfação laboral tem desde então vindo a ser explorado por diversas áreas, tais como a psicologia (Locke, 1976), sociologia (Peter Warr, 1975) e economia (Bakotic, 2016). Um dos motivos pelos quais persiste o estudo da satisfação laboral é o facto da sociedade estar em constante evolução, e assim sendo, o que proporcionava uma maior satisfação laboral há 50 anos pode diferir dos dias de hoje.

## 2. Objetivos

Este estudo tem como principal objetivo explicar o que leva a um individuo sentir-se satisfeito com o seu trabalho em geral. Como variáveis preditoras temos variáveis de caracterização (Género, Classe etária, Habilitações literárias, Antiguidade na organização e Posição hierárquica) e os indicadores Infraestrutura Laboral, Envolvimento organizacional, Orientação para o cliente, Relacionamentos interpessoais, Autonomia, Reconhecimento, Desenvolvimento profissional e Remuneração/ benefícios.

Para além de estudar a satisfação laboral, o estudo também tem como objetivo estudar os indicadores através das variáveis de caracterização.

### 3. Resumo Teórico

Neste capítulo irá ser feita uma breve abordagem às metodologias estatísticas utilizadas neste projeto.

A não demonstração de alguns resultados e a não exaustividade na descrição dos mesmos deve-se ao facto de já terem sido lecionados na licenciatura e/ ou no presente mestrado.

#### 3.1 Da Regressão Linear Múltipla ao Modelo Linear Generalizado

Em estatística, na área denominada por Regressão, tem-se como objetivo explorar o relacionamento ou influência de um conjunto de variáveis  $X$ , (denominadas por variáveis explicativas, ou preditores) com outra variável designada como dependente ou resposta ( $Y$ ).

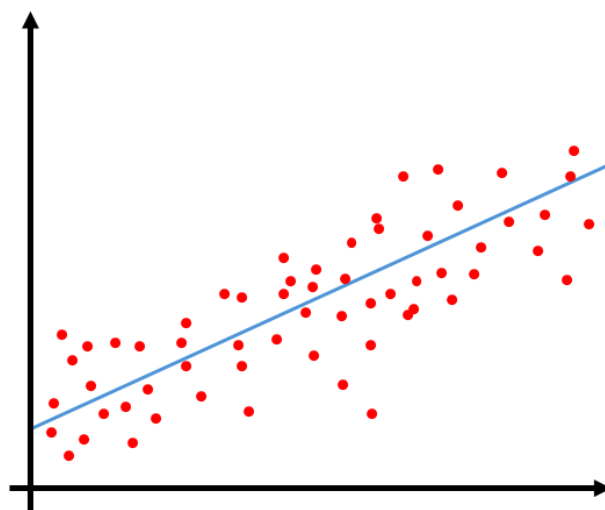
O aproveitamento desse relacionamento é traduzido por um modelo matemático em que a variável dependente é a soma de uma componente aleatória de valor médio nulo e de uma componente determinística, que é função dos preditores, e que traduzirá o valor médio da variável dependente condicional a um conjunto de valores fixos dos preditores.

Inicia-se a abordagem pelo modelo de regressão linear simples, pela sua maior simplicidade e porque historicamente foi o primeiro.

#### 3.2 Regressão linear simples

Como foi dito anteriormente, as relações entre variáveis não são deterministas, ou seja, não existem fórmulas matemáticas que relacionem variáveis. Por mais que se tente controlar o valor de fatores que determinam uma variável, existe sempre uma componente aleatória na resposta.

O modelo de regressão linear simples (MRLS) é inspirado em diagramas de dispersão em que os pontos se “espalham” ao longo de uma reta, ou seja, grandezas cujo ritmo de variação da variável resposta é aproximadamente igual, independentemente do nível do preditor.



A reta traçada a azul idealiza o valor médio condicional da variável resposta, mas o que é observado são os pontos a vermelho. Ao realizar um grande número de observações para uma determinada abscissa fixa, o valor médio corresponderá ao valor da reta para essa abscissa. O modelo adotado será então:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

em que  $\beta_0, \beta_1$  são parâmetros desconhecidos (ordenada na origem e declive da recta), os valores  $(x_1, \dots, x_n)$  são constantes conhecidas e  $(\varepsilon_1, \dots, \varepsilon_n)$  é um vetor estocástico i.i.d. de valor médio nulo e homocedástico. Como foi dito anteriormente, para se poder fazer inferência sobre os parâmetros é usual assumir a gaussianidade dos erros.

Utiliza-se usualmente o método dos mínimos quadrados para estimar os parâmetros da reta. Denominando a soma de quadrados por SQ,

$$SQ(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \times x_i)^2.$$

Assim, para encontrar os minimizantes da função anterior, é necessário resolver o sistema de equações normais com as derivadas parciais de 1ª ordem igualadas a zero. Depois de obtidos os candidatos a extremantes, é necessário garantir que o determinante da matriz hessiana nos mesmos é positivo. Não são incluídos aqui os cálculos pelos mesmos constarem em qualquer manual de regressão. Chega-se assim aos estimadores:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{\text{cov}(x,y)}{s_x^2} = r \times \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

A reta estimada passa pelo centro de gravidade bidimensional da nuvem de pontos  $(\bar{x}, \bar{y})$ , basta olhar para o estimador da ordenada na origem da reta para verificar este facto.

Aplicando as propriedades do valor médio, variância e covariância e atendendo aos pressupostos do modelo é possível provar que estes estimadores são centrados e que:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Repare-se que, quanto maior a variabilidade das observações em torno da reta (traduzida pelo  $\sigma^2$ ), maior a incerteza das nossas estimativas, como seria de esperar. Isto é um fator que não se controla, faz parte da aleatoriedade do sistema.

No entanto, quanto maior a dimensão da amostra e maior a variabilidade do preditor, menor a incerteza das nossas estimativas.

É possível provar também que  $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \times \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ .

Mais uma vez, observa-se que os mesmos fatores que aumentam a variância dos estimadores, aumentam a correlação linear entre as duas estimativas em magnitude e que o sinal desta é oposto à média do preditor.

Sob a hipótese de dados gaussianos, é possível provar que estes estimadores são de variância mínima (BLUE- *best linear unbiased estimator*), ou seja, são estimadores lineares e a sua variância iguala o LIFCR (limite inferior de Fréchet-Cramer-Rao).

Estimada a reta, podem ser feitas previsões para a resposta face a um certo valor do preditor,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i.$$

É chamado de resíduo à estimativa que é dada para o valor assumido pela variável  $\varepsilon_i$ ,  $r_i = y_i - \hat{y}_i$ . Pode-se provar que a sua soma é nula, basta atender à expressão dos estimadores.

O método dos mínimos quadrados não produz um estimador para o terceiro parâmetro do modelo,  $\sigma^2$ . Repare-se que, sem estimar este parâmetro desconhecemos as variâncias dos outros dois estimadores e isso impossibilita fazer inferência para os mesmos.

Dado que a componente errática tem média nula, a sua variância coincide com o valor médio do erro ao quadrado. Isso sugere que o estimador da variância é a média das estimativas dos quadrados dos erros, ou seja,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n}.$$

É possível estabelecer duas relações lineares entre os resíduos, ou seja, é possível exprimir dois resíduos como combinação linear dos restantes. Assim, este estimador é enviesado pois na realidade está a somar o quadrado de  $n - 2$  resíduos. Então, um estimador centrado da variância dos resíduos será

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-2}.$$

Se não for assumida uma distribuição para os erros do modelo, mais nada poderá ser dito sobre o mesmo. Se se assumir a hipótese de v.a.'s i.i.d. gaussianas, outras janelas se abrirão.

Como os estimadores dos parâmetros da reta são combinações lineares de variáveis aleatórias independentes com distribuição normal, aplicando o Teorema da Aditividade da Normal deduz-se que:

$$\hat{\beta}_0 \cap N\left(\beta_0, \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right),$$

$$\hat{\beta}_1 \cap N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Ainda não se tem uma variável fulcral porque as distribuições anteriores dependem do parâmetro desconhecido  $\sigma^2$ . É necessário studentizar as estatísticas anteriores.

Demonstra-se a independência entre  $\hat{\sigma}^2$  e  $(\hat{\beta}_0, \hat{\beta}_1)$ . Acaba por ser consequência da propriedade, que apenas se verifica no modelo gaussiano do estimador da média ( $\bar{x}$ ), ser independente do estimador da variância ( $s^2$ ). Sob a validade deste modelo, além da independência, deduz-se que:

$$T = (n - 1) \times \frac{\hat{\sigma}^2}{\sigma^2} \cap \chi_{n-2}^2,$$

e isto permite efetuar a studentização, provando-se que:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2},$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}.$$

Atendendo às características das variáveis anteriores ((1) a sua distribuição não depender de parâmetros desconhecidos, (2) o único parâmetro desconhecido ser o parâmetro da reta e (3) como função deste ser uma função monótona), poderão ser usadas como variáveis fulcrais na construção de estimadores intervalares para os parâmetros e como estatísticas de testes paramétricos.

Assim,

$$IC_{1-\alpha}(\beta_i) = ]\hat{\beta}_i \mp t_{n-2; 1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\beta}_i}[ , i = 1, 2.$$

Adquire particular relevância o teste:

$$H_0: \beta_1 = 0,$$

pois a sua não rejeição conduz à não aceitação do modelo de previsão, dado que as observações serão meras flutuações em torno de um valor médio qualquer que seja o valor do preditor. Repare-se que, isto acontecerá para valores muito baixos de  $\hat{\beta}_1$ , pelo menos se comparados com a estimativa da sua incerteza.

Este estimador é diretamente proporcional à correlação empírica entre as variáveis, pelo que a não rejeição de  $H_0$  se verificará em ambientes de correlação muito fraca.

O teste  $t$  anterior poderá ser substituído por um teste equivalente usando a distribuição  $F$ , em que a estatística de teste é o quociente entre os quadrados médios da regressão e os quadrados médios dos erros. Isto é, a variabilidade total da variável resposta ( $SQTot = (n - 1)s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ ) é decomposta em 2 fontes, uma explicada pelo modelo ( $SQReg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ ) e outra residual (não explicada pelo modelo,  $SQRes = \sum_{i=1}^n (\hat{y}_i - y_i)^2$ ). O quociente entre os valores médios dessas variabilidades origina a estatística do teste.

Na regressão múltipla existem diferenças entre o teste  $t$  e o teste  $F$ , sendo este um teste global e o primeiro um teste local. Na regressão simples os dois testes são idênticos.

Tabela 1 - Tabela Anova

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Regressão	1	$SQReg$	$SQReg = MSReg$	$F = \frac{MSReg}{MSRes}$
Resíduo	$n - 2$	$SQRes$	$\frac{SQRes}{n - 2} = MSRes$	
Total	$n - 1$	$SQTot$	$\frac{SQTot}{n - 1} = MSTot$	

Uma forma de avaliarmos a adequabilidade do modelo é atendermos ao coeficiente de determinação, que se obtém levantando ao quadrado o coeficiente de correlação:

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

O  $R^2$  varia entre 0 e 1. Valores próximos de zero revelam uma total incapacidade do modelo de explicação das variações da variável resposta face às homólogas do preditor.

No fim, deve-se verificar se os pressupostos que fundamentaram a análise, nomeadamente, a normalidade, homocedasticidade e a independência dos resíduos entre si, têm consistência com o que foi observado. Para isso efetuam-se testes aos resíduos do modelo, que são estimativas dos valores dos erros.

Os métodos de estimação, os testes e intervalos de confiança que se apresentam para o modelo linear são bastante robustos, relativamente ao pressuposto da normalidade, que é uma consequência do Teorema do Limite Central e suas generalizações para variáveis não identicamente distribuídas e da Lei Fraca dos Grandes Números.

No entanto, em certos casos, a distribuição de probabilidade da variável dependente afasta-se de tal modo da distribuição normal, que é necessário desenvolver novos modelos de regressão baseados em pressupostos diferentes.

## 3.3 Regressão linear múltipla

### 3.3.1 Estimadores dos parâmetros e as suas propriedades

No modelo de regressão linear múltipla (MRLM) clássico,

$$Y = X\beta + \varepsilon,$$

onde  $X$  é uma matriz de dimensão  $n \times (k + 1)$ , ( $n > k + 1$ ) com a primeira coluna constituída pela unidade e as restantes pelos valores das covariáveis para cada um dos  $n$  registos independentes da amostra.  $\varepsilon$  é uma matriz de dimensão  $n \times 1$  constituída por  $n$  variáveis aleatórias i.i.d. de valor médio nulo. Com o objetivo de podermos fazer inferência estatística sobre os parâmetros desconhecidos do modelo, é vulgar assumirmos a distribuição normal para a sua modelação.  $\beta$  é um vetor de parâmetros desconhecidos também de dimensão  $(k + 1) \times 1$ ,  $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$ , em que  $\beta_0$  traduz a componente fixa e os restantes traduzem o coeficiente associado a cada um dos preditores.

Logo,

$$\forall i \in \{1, 2, \dots, n\}, y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i.$$

O método dos *MQ* (mínimos quadrados) escolhe as estimativas dos parâmetros desconhecidos que minimizam a soma dos quadrados dos resíduos da regressão.

$$\varepsilon^T \times \varepsilon = (Y - X\beta)^T \times (Y - X\beta) = 0.$$

Assim, o sistema de equações normais consiste em igualar a zero as derivadas parciais em ordem aos parâmetros. Assumindo que a matriz  $(X^T X)$  é uma matriz regular, a solução deste sistema é dada por:

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

Note-se que se o determinante da matriz  $X^T X$  estiver próximo de zero, o que acontecerá se existirem preditores que são aproximadamente combinações lineares dos restantes, existem problemas de instabilidade na sua matriz inversa. Assim, pequenos erros de arredondamento nos valores dos preditores poderão provocar alterações drásticas nas estimativas dos parâmetros.

De forma análoga à regressão linear simples, teremos que:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY \text{ em que } H = X(X^T X)^{-1} X^T,$$

denominada de matriz hat, e o vetor de resíduos é:

$$R = \hat{\varepsilon} = Y - \hat{Y}.$$

É possível provar que  $H$  é uma matriz simétrica e idempotente e que:

1. A soma dos resíduos é nula.
2. A média da variável observada é igual à média das previsões para a mesma.
3. A covariância empírica entre cada preditor e o vetor de resíduos é nula.

Cada elemento de  $\hat{\beta}$  é uma estimativa da variação da variável resposta por aumento unitário do preditor, mantendo-se os níveis dos outros preditores constantes.

Prova-se que a matriz de covariâncias de  $\hat{\beta}$  é  $\text{cov}(\hat{\beta}) = \hat{\sigma}^2 \times (\mathbf{X}^T \mathbf{X})^{-1}$  (o estimador de  $\sigma^2$  será definido mais à frente).

Prova-se também que estes estimadores são centrados e de variância mínima na classe dos estimadores lineares e que  $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2 \times (1 - R_j^2)}$ , em que  $R_j^2$  é o coeficiente de determinação da regressão múltipla de  $X_j$ , usando como preditores todos os outros preditores com exceção dele próprio.

Podemos fazer considerações semelhantes às que foram feitas na RLS:

1. Quanto maior  $\sigma^2$ , ou seja, quanto maior for a variabilidade da componente errática, maior a incerteza nas previsões.
2. Quanto maior a dimensão amostral e quanto maior for a variabilidade dos preditores maior a certeza nas previsões.
3. A existência de relações lineares significativas entre os preditores é uma situação perigosa, pois aumenta a variabilidade das estimativas, podendo causar instabilidade no modelo. Isto está relacionado com o facto do determinante da matriz de hat estar próximo de zero e consequentemente aumentar a instabilidade da sua inversa.

### 3.3.2 Inferência

Pelas mesmas razões referenciadas no MRLS, os estimadores dos parâmetros são combinações lineares de normais independentes, e atendendo ao Teorema da Aditividade da Normal:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\beta_j}} \cap N(0,1).$$

Mas dado que o desvio padrão do estimador depende da variância dos erros, parâmetro também desconhecido, esta variável não poderá ser variável fulcral para fazer inferência sobre  $\beta_j, j = 0, \dots, k + 1$ .

A estatística terá de ser studentizada, da mesma forma que se fez no modelo MRLS.

Assim, é usada a soma dos quadrados dos resíduos ponderada pelos graus de liberdade para estimar a variância comum dos erros. Como a soma dos resíduos é nula, a sua variância será estimada pela média dos quadrados dos resíduos. Por outro lado, como os estimadores dos parâmetros são combinações lineares das observações, é possível estabelecer uma relação entre  $k + 1$  resíduos e os restantes, pelo que só  $n - (k + 1)$  resíduos são independentes:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n - k - 1}.$$

Assim,

$$\widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2 \times (1 - R_j^2)}$$

e da mesma forma que no MRLS,

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \cap t_{n-k-1},$$

será a variável fulcral que permitirá fazer inferência sobre cada  $\beta_j$  individualmente.

Assim,

$$IC_{1-\alpha}(\beta_j) = ]\hat{\beta}_j \mp t_{n-k-1; 1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\beta}_j}[, \quad j = 0, \dots, k.$$

É possível também estabelecer intervalos de confiança para combinações lineares dos parâmetros, como, por exemplo, o intervalo de confiança para o valor médio da variável resposta, condicional a um vetor fixo de valores dos preditores,  $\mathbf{X}_0$ :

$$IC_{1-\alpha} \left( \beta_0 + \sum_{j=1}^k \beta_j \times X_{0j} \right) = ]\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j \times X_{0j} \mp t_{n-k-1; 1-\frac{\alpha}{2}} \times \sqrt{\hat{\sigma}^2 \times \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_0}].$$

### 3.3.3 Medidas de ajustamento e teste sobre a significância global do modelo

A mesma relação, anteriormente citada na regressão linear simples, verifica-se:

$$\begin{aligned} SQTot &= SQRes + SQReg \Leftrightarrow \\ \Leftrightarrow ((n-1)s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2) &= \sum_{i=1}^n (\hat{y}_i - y_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \end{aligned}$$

O coeficiente de determinação  $R^2 = \frac{SQReg}{SQT}$  determina a fração de variabilidade da resposta que é explicada pelo modelo. Quanto mais próximo de 1, maior o ajustamento do mesmo à amostra multivariada.

Verificamos que o  $R^2$  nunca decresce, e, em geral, cresce quando adicionamos uma outra variável explicativa na regressão. Isso porque, a soma dos quadrados dos resíduos nunca aumenta quando adicionamos uma nova variável explicativa na regressão. Torna-se, portanto, um pouco mais difícil decidir sobre a inclusão ou não de uma variável adicional no modelo. Na verdade, desejar-se-ia saber se essa variável tem um efeito pouco significativo na explicação de  $Y$ . Muitas vezes, é interessante utilizar o conceito do  $R^2$  ajustado, que irá considerar o número de variáveis explicativas no modelo, ou seja:

$$\bar{R}^2 = 1 - \frac{\frac{SQRes}{n-k-1}}{\frac{SQTot}{n-1}} = 1 - \frac{\hat{\sigma}^2}{\frac{SQTot}{n-1}} = 1 - (1 - R^2) \times \frac{n-1}{n-k}.$$

Este coeficiente poderá ser negativo e a introdução de um novo preditor não incrementa necessariamente o mesmo, pois penaliza também a entrada de novos preditores.

Estes coeficientes são medidas de ajustamento do modelo, mas não dão uma resposta à significância global do modelo.

Muitas vezes pretende-se testar se os coeficientes são conjuntamente iguais a zero, ou seja, testar a significância global do modelo:

$$H_0: \beta_1 = \dots = \beta_k = 0$$

$$H_1: \exists i \in \{1, 2, \dots, k\}: \beta_i \neq 0$$

O teste pode ser generalizado. Se considerarmos  $A$  um subconjunto estrito de preditores:

$$H_0: \beta_j = 0, \forall j \in A$$

$$H_1: \exists j \in A: \beta_j \neq 0$$

Se chamarmos modelo restrito ao modelo com todos os preditores com exceção dos que constam em  $A$ ,  $H_0$  defende que devemos rejeitar o modelo global em detrimento do modelo restrito, modelo sem os preditores pertencentes a  $A$ . O primeiro caso abordado coincide com o caso particular do modelo restrito ser o modelo nulo, só com a constante, sem preditores.

É possível provar que a estatística:

$$F = \frac{\frac{SQReg_{restrito} - SQReg_{\text{n\~{a}o restrito}}}{q}}{\frac{SQRes_{\text{n\~{a}o restrito}}}{n-k-1}} = \frac{\frac{R^2_{\text{n\~{a}o restrito}} - R^2_{restrito}}{q}}{\frac{1 - R^2_{\text{n\~{a}o restrito}}}{n-k-1}} \cap F_{q, n-k-1},$$

sendo  $q$  o cardinal do conjunto  $A$ .

Se a estatística  $F$ , sendo sempre positiva, tomar um valor muito alto, isto significa que a soma dos quadrados dos erros do modelo restrito é significativamente maior do que a do modelo global, devendo rejeitar  $H_0$ . O ponto crítico será um quantil da distribuição  $F$  de ordem  $1 - \alpha$ .

### 3.3.4 Escolha do modelo

Antes de definir o teste para comparação de modelos, é necessário introduzir alguns conceitos.

O número de modelos possíveis é determinado por  $2^k$ , sendo  $k$  o número de variáveis preditoras. Neste contexto, existem diversos modelos, variando o grau de explicação e parcimónia, oscilando ente dois extremos:

- Modelo completo ou saturado, também dito de maximal, que inclui todas as covariáveis em estudo. Trata-se de um modelo que tem a máxima “adaptação” aos dados, mas poderá ser extremamente “complicado”, pois pode incluir variáveis em número excessivo, algumas pouco

ou nada explicativas, outras cuja influência na variável resposta pode ser substituída por outro preditor;

- Modelo nulo, também dito de minimal, como sendo o modelo que contém apenas uma constante, ou seja, em que todas as variáveis  $Y_i$  têm o mesmo valor médio,  $\mu$ , pois não existem variáveis explicativas. Trata-se de um modelo pouco interessante, uma vez que o modelo irá prever todos os casos com a média empírica da variável resposta.

Será chamado de modelo corrente aquele que estará entre os dois, incluindo alguns preditores e excluindo outros.

Existem vários métodos para a escolha do modelo ótimo, como o *forward*, *backward*, *stepwise*, com ou sem efeitos de interação entre as variáveis, com opções de escolha nos critérios de entrada ou saída das variáveis, *etc.* Uma consulta aos menus do SPSS ou do R rapidamente nos ilustram alguns deles. No método *forward*, o ponto de partida é o modelo nulo. Verifica-se, usando um certo critério, se existe algum preditor fora do modelo que justifique a sua inclusão. Se não existir, o modelo é encerrado, se existir, inclui-se e refaz-se o modelo, voltando a considerar se existe alguma variável de fora que valha a pena incluir. No método *backward*, o ponto de partida é o modelo maximal e vai-se excluindo preditores enquanto houver justificção para tal. A variante *stepwise*, por sua vez, combina os dois métodos.

Observa-se então, que existem vários caminhos para a escolha do modelo, que não originam o mesmo necessariamente.

Em geral é usado o critério AIC (*Akaike's Information Criterion*) como critério de seleção. Assim, após a aplicação do método escolhido, o modelo que apresentar um menor AIC será selecionado. O critério AIC é definido da seguinte forma:

$$AIC = -2\ln(L(\hat{\beta})) + 2q,$$

em que  $\ln(L(\hat{\beta}))$  representa a função log-verossimilhança para o modelo com  $q$  variáveis explicativas que adotamos (logo  $q + 1$  parâmetros). A estatística que compara dois modelos não se pode basear apenas na qualidade do ajustamento, pois sendo assim, o melhor modelo seria o maximal. Tem de penalizar a entrada de um preditor, se a sua entrada significar uma melhoria pouco significativa do ajustamento.

### 3.3.5 Exigências e recomendações do modelo de regressão linear múltipla

Além das exigências de normalidade, independência e homocedasticidade para a variável dependente, o que implicará uma análise dos resíduos do modelo adequada no sentido de testar a veracidade destes pressupostos, antes de implementar o MRLM deve-se verificar se certas condições se verificam.

#### Normalidade

Para testar a normalidade dos resíduos temos os testes de Shapiro-Wilk, Kolmogorov-Smirnov, ajustamento do qui-quadrado, entre outros. Qualquer um destes poderá ser aplicado aos resíduos para averiguar a bondade do pressuposto bem como métodos gráficos, como por exemplo o Q-Q plot.

#### Homocedasticidade

Para testar a homocedasticidade dos resíduos existem variadíssimos testes, tais como o teste de White, o teste de Goldfeld-Quandt, bem como abordagens gráficas, traçando o gráfico dos resíduos padronizados *versus* valores ajustados e verificando a existência de padrões.

## Independência

Para testar a independência dos resíduos existem, por exemplo, os testes de Durbin-Watson e o de Ljung-Box. Adicionalmente, existem ainda métodos gráficos com representação da função de autocorrelação e autocorrelação parcial dos resíduos.

Este não é o espaço mais adequado para descrever estes testes, mas é possível encontrar informação sobre os mesmos na bibliografia relacionada com o tema.

É possível realizar estes testes com recurso a pacotes estatísticos como o R, Python e SPSS. É sempre recomendável combinar estes testes com a análise visual dos resíduos para uma avaliação mais completa.

## Multicolinearidade

Um dos problemas que surge frequentemente quando se usam muitos preditores em conjunto é a sua não independência. Podem existir variáveis independentes que estejam ligadas de tal forma que seja muito difícil impedir que os seus efeitos não se confundam entre si potenciando-se ou anulando-se.

Este problema resolve-se eliminando algumas das variáveis cujo papel influenciador na variável dependente já é efetuado por outra covariável, ou então faz-se uma análise de componentes principais e substitui-se os preditores originais pelas suas componentes principais. Neste caso, pode haver problemas na sua interpretabilidade pois as novas variáveis independentes são combinações das originais.

A presença de covariáveis que são aproximadamente combinações lineares de outras variáveis leva a que exista correlação elevada entre elas, e assim a interpretação de quais as variáveis a incluir no modelo pode ficar comprometida.

Assim, para a análise de multicolinearidade deve se considerar o cálculo do *VIF* - *Variance Inflation Factor*, que indica a existência ou não de multicolinearidade, sendo dado por:

$$VIF = \frac{1}{1-R^2},$$

em que  $R^2$  é o coeficiente de determinação da regressão de uma covariável em relação às restantes variáveis do modelo.

No caso de existirem covariáveis com VIF elevado ( $VIF > 10$ ), a solução passa por retirar uma dessas covariáveis e refazer a análise.

## Interação

Existe efeito de interação quando a força da associação entre a variável independente e a variável dependente é diferente para diferentes níveis da variável independente.

Os diferentes *softwares* permitem a introdução automática dos efeitos de interação entre todos os preditores, mas isso torna o modelo mais complexo. Por isso, apenas com razões muito fortes se deve recorrer à introdução de novas variáveis que reflitam este facto.

## Observações influentes

Uma observação é considerada influente se a sua exclusão produz alterações significativas nas estimativas dos parâmetros do modelo. A sua presença pode assim originar conclusões que não correspondem à realidade.

As formas mais comuns de deteção destas observações influentes é através das medidas de Cook ou medidas *leverage*.

A distância de Cook é uma medida de influência baseada no valor estimado dos coeficientes da regressão. Para o seu cálculo deve considerar-se a amostra sem a observação que se pensa ser influente

e a amostra com esta observação. Se o valor estimado “se alterar muito”, considera-se que a observação pode ser influente. O cálculo da distância de Cook é dado por:

$$\Delta\hat{\beta}_j = (\hat{\beta} - \hat{\beta}_{(-j)})^T (X^T Y X) (\hat{\beta} - \hat{\beta}_{(-j)}),$$

em que  $\hat{\beta}_{(-j)}$  representa as estimativas dos parâmetros do modelo retirando a  $j$ -ésima observação que queremos verificar se é influente ou não.

A análise dos resíduos é importante para avaliar os *outliers* que podem comprometer o desempenho do modelo. Todos os estimadores de regressão são baseados em médias e, portanto, têm limite de rutura nulo, ou seja, basta uma observação tender para infinito para eles também terem esse comportamento.

Existem várias medidas aplicadas dentro da Regressão Linear para avaliar esses casos influentes, todos baseados nos resíduos do modelo - os valores de “*leverage*” (valores de alavanca que traduzem a forma como podem influenciar uma estimativa de um parâmetro).

Neste projeto, o foco será dado aos resíduos de Pearson ou também chamados de resíduos standardizados e à Distância de Cook para avaliar o comportamento dos *outliers*.

Denominando como  $rp_j$ - resíduo de Pearson da  $j$ -ésima observação, vem que,

$$rp_j = \frac{r_j}{\hat{\sigma}}.$$

São uma versão standardizada dos resíduos, pois têm média nula e desvio padrão unitário.

Define-se “*leverage*” de uma observação,

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, j = 1, \dots, n.$$

Os valores de “*leverage*” ou valores de alavanca variam entre  $\frac{1}{n}$  e 1. E quanto maior for o seu valor maior o impacto nas estimativas dos coeficientes do modelo. É útil observar gráficos que representam os valores de alavanca *versus* os resíduos padronizados.

### 3.3.6 Modelo linear generalizado

O pressuposto de linearidade entre a variável resposta e as variáveis explicativas, e a normalidade no modelo de regressão linear, premissa necessária para testar hipóteses e construir intervalos de confiança para os parâmetros, é uma condição que exclui uma gama de situações como por exemplo variáveis binárias, de proporção, de contagem ou variáveis categóricas. Apesar de serem ainda bastante utilizados pela comunidade acadêmica até aos dias atuais, os MRLM são desajustados para uma série de situações, e modelam somente observações com distribuição normal.

Como forma de ultrapassar as limitações nos MRLM, existem propostas de novas classes de modelos lineares generalizados (MLG), com o propósito de dar uma série de opções para a distribuição da variável resposta para além da normal. Os MLG, tornam flexível a relação da variável resposta com as covariáveis por meio de uma função de ligação ( $g(\cdot)$ ), que tem por objetivo garantir que os parâmetros possam assumir quaisquer valores reais, preservando os valores ajustados no domínio da distribuição.

Modela-se também o valor esperado da variável resposta, e incluem o MRLM se considerarmos que a função de ligação é a função identidade. Saem fora do âmbito desta tese, pelo que é apenas citado a sua existência.

### 3.4 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é uma extensão do teste de Mann-Whitney em que se pretende decidir sobre a hipótese de igualdade das localizações centrais de  $k$  populações independentes, sem se admitir o pressuposto de dados gaussianos.

Combinam-se as  $k$  amostras numa só e atribui-se a cada elemento desta amostra,  $x_{ij}$ , o seu *rank*,  $r(x_{ij})$ . O primeiro índice indica a população de onde foi recolhido esse elemento e o segundo a ordenação da observação dentro da amostra desta população.

Soma-se, a seguir, os *ranks* de todas as observações de cada uma das subamostras,

$$R_i = \sum r(x_{ij}), i = 1, \dots, k.$$

O que se pretende é fazer uma análise de variância destes *ranks*, verificando se o *rank* médio das subamostras difere do *rank* médio global,  $\frac{N+1}{2}$ , sendo  $N$  o número total de observações da amostra combinada.

Este afastamento quadrático terá de ser ponderado em cada amostra pela sua dimensão. Então, toma-se como medida da discrepância global,

$$D = \sum_{i=1}^k n_i \times \left( \frac{R_i}{n_i} - \frac{N+1}{2} \right)^2 = \sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)}{4}.$$

Sob a validade de  $H_0$  é possível provar que  $E(D) = \frac{N(N+1)(K-1)}{12}$  e a estatística de teste de Kruskal-Wallis será uma mudança de escala da estatística  $D$ , ou seja,

$$H = \frac{D}{\frac{N(N+1)}{12}},$$

Pretende-se que sob validade da hipótese nula o valor médio da estatística de teste não dependa do número de observações, mas sim apenas do número de grupos envolvidos, neste caso,  $E(H) = k - 1$ .

A estatística  $H$ , sendo uma soma de quadrados de variáveis, que, devido ao Teorema de Limite Central, podem ser considerados aproximadamente gaussianos, tem uma distribuição que pode ser bem aproximada por uma qui-quadrado com  $k - 1$  graus de liberdade. Para alguns valores de  $N$  existem tabelas de pontos críticos e probabilidades que lhes estão associadas. Estas tabelas foram construídas deduzindo a distribuição exata de  $H$  sob a veracidade de  $H_0$ .

Em toda a descrição anterior, parte-se do princípio de que não existem empates nas observações, o que é uma suposição perfeitamente válida para variáveis absolutamente contínuas. No entanto, a discretização das mesmas na sua mensuração, torna possível a existência de empates. Nestas situações

é vulgar a atribuição de *ranks* médios às observações empatadas e a correção da estatística  $H$  para empates.

### 3.4.1 Teste de post hoc de Dunn

O teste de comparações múltiplas de Dunn é utilizado após o teste de Kruskal-Wallis, se e somente se é rejeitada a hipótese nula. Por este motivo, é chamado muitas vezes de post hoc de Dunn. A alternativa a este teste seria realizar uma multiplicidade de testes de Mann-Whitney comparando as localizações de todos os pares de grupos. Só que para que o nível de significância global se mantenha, obriga a que se realize cada teste de comparação de Mann-Whitney a um nível de significância muito baixo o que leva quase de certeza à aceitação da igualdade dos parâmetros de localização.

A estatística do teste Dunn é baseada na diferença entre os postos médios dos grupos comparados, ajustada por um fator de correção,

$$Z = \frac{|R_i - R_j|}{\sqrt{\frac{k(k+1)}{12N}}}$$

Tal como no teste de Kruskal-Wallis, a região de rejeição é unilateral direita. Não existem tabelas publicadas com os pontos críticos deste teste para dimensões amostrais pequenas, mas para dimensões grandes é possível provar a normalidade da estatística de teste e usar os pontos críticos da normal *standard*.

Detetam-se assim quais os grupos responsáveis pela rejeição da hipótese nula no teste de Kruskal-Wallis.

## 3.5 Coeficiente alfa de Cronbach

O alfa de Cronbach é uma medida de confiabilidade usada em estatística para avaliar a consistência interna de um conjunto de itens em um teste ou questionário. Foi desenvolvido por Lee Cronbach em 1951.

$$\alpha = \frac{k}{k-1} \times \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_T^2}\right),$$

$K$  - Número de grupos;

$\sigma_i^2$  - Variância dentro do grupo  $i$ ;

$\sigma_T^2$  - Variância dentro de todos os grupos.

Ele varia entre 0 e 1, se está próximo de 1 indica uma alta consistência interna entre os itens, sugerindo que eles medem a mesma construção ou traço latente. Se está próximo de zero, sugere uma fraca consistência interna, indicando que os itens não estão fortemente correlacionados e podem não medir o mesmo construto. No entanto, é importante notar que o alfa de Cronbach assume a unidimensionalidade da escala, ou seja, que todos os itens medem uma única característica subjacente. Se esta suposição não for atendida, outras técnicas podem ser mais apropriadas.

Usualmente, valores do alfa de Cronbach entre 0.7 e 0.9 são considerados aceitáveis e indicam uma consistência interna moderada a alta.

## 4. Metodologia

Este capítulo tem como objetivo dar uma visão abrangente das abordagens e dos procedimentos que serão empregues, desde o momento da construção do questionário até à demonstração dos resultados.

### 4.1 Questionário

O presente estudo inicia-se com a elaboração do questionário. O mesmo determina a qualidade dos dados e consequentemente a qualidade do *output* final. Assim sendo, é necessário seguir um conjunto de regras e boas práticas no momento da construção do mesmo.

Ele foi elaborado pela consultora Qmetrcis e possui carácter confidencial. Ainda assim, de forma a envolver o leitor, apresenta-se de seguida a sua estrutura e algumas das boas práticas aplicadas ao mesmo no momento da sua construção.

#### **Estrutura do questionário:**

O questionário é composto por 57 questões. As cinco primeiras visam identificar os colaboradores quanto à Posição hierárquica, Classe etária, Antiguidade na organização, Habilitações literárias e Género. As restantes são questões relacionadas com o contexto organizacional, sendo medidas na escala de Likert onde 1 representa o nível mais baixo e 10 o nível mais alto de satisfação/ concordância.

#### **Boas práticas aplicadas ao questionário:**

**Contextualização** – Contextualizar os colaboradores acerca da finalidade do estudo é crucial, especialmente num estudo que visa melhorar a satisfação laboral.

**Anonimato** – O anonimato assume um papel fundamental já que sem a garantia do mesmo existe uma maior propensão de respostas desonestas. Para além de ser mencionado no início do questionário o seu anonimato, ele é realizado em formato papel, o que reforça ainda mais a segurança sentida pelos colaboradores.

**Linguagem simples e específica** – Uma linguagem simples e específica não só reduz a hipótese de ambiguidades, como também reduz o esforço necessário para a interpretação das questões. Questionários em que é necessário um maior esforço de interpretação podem deteriorar a qualidade das respostas e até mesmo levar à desistência dos inquiridos.

**Sequência de perguntas:** A sequência de perguntas é clara e fluída, permitindo uma fácil compreensão por parte dos colaboradores. É importante evitar mudanças repentinas entre temas pois isso pode causar cansaço, que, como mencionado no parágrafo anterior, pode deteriorar a qualidade das repostas e inclusive levar à desistência dos inquiridos.

## 4.2 Procedimento

Após a obtenção da base de dados proveniente do questionário, agrupou-se 40 questões em 8 indicadores por meio de uma análise de consistência interna, utilizando o coeficiente alfa de Cronbach.

Foi feita uma análise descritiva à população em estudo, aos 8 indicadores criados e à satisfação global.

Recorrendo aos testes de Mann-Whitney, Kruskal-Wallis e Dunn, fez-se uma análise comparativa dos indicadores mais a satisfação global por todas as variáveis de caracterização.

Por último, construiu-se um modelo de regressão linear múltipla com o objetivo de explicar a satisfação global, sendo avaliado a qualidade do mesmo.

As ferramentas usadas neste projeto foram o Excel e SPSS.

## 5. Resultados

### 5.1 Análise exploratória dos dados

Antes de construir o modelo de regressão linear múltipla, apresentam-se as variáveis intervenientes.

A nossa variável dependente, satisfação global com a empresa (SGlob), foi medida na escala de Likert em que 1 é o nível de satisfação mais baixo e 10 o nível de satisfação mais alto.

Dentro das variáveis explicativas temos variáveis de caracterização e variáveis que são a média ponderada de outras variáveis medidas na escala de Likert, chamemos a essas de indicadores.

#### 5.1.1 Variáveis de caracterização

As variáveis de caracterização são o Género, Classe etária, Habilitações literárias, Antiguidade na organização e Posição hierárquica.

Aproximadamente dois terços dos colaboradores são mulheres. As classes etárias mais frequentes são 26-35 e 36-45 anos. Apenas 8.9% têm grau de ensino superior, sendo o 3º ciclo e o ensino secundário as categorias mais frequentes, ambas com proporções semelhantes. Cerca de 45% dos colaboradores se encontram na organização há menos de 5 anos. Por fim, mais de três quartos da população são operadores/ vendedores.

Tabela 2 - Caracterização dos colaboradores

Variável	Categorias	Frequência	Percentagem	Variável	Categorias	Frequência	Percentagem
<b>Género</b> C5	1. Masculino	271	30,3%	<b>Antiguidade na organização</b> C3	1. < 1 ano	169	18,9%
	2. Feminino	557	62,2%		2. 1-5 anos	240	26,8%
	NS\ NR	57	6,4%		3. 6-10 anos	71	7,9%
<b>Classe etária</b> C2	1. 18-25 anos	148	16,5%		4. 11-20 anos	210	23,5%
	2. 26-35 anos	274	30,6%		5. > 20 anos	174	19,4%
	3. 36-45 anos	302	33,7%	NS\ NR	21	2,3%	
	4. 46-55 anos	118	13,2%	<b>Posição hierárquica</b> C1	1. Operador/ Vendedor	676	75,5%
	5. +55 anos	21	2,3%		2. Supervisor	32	3,6%
NS\ NR	22	2,5%	3. Coordenador equipa/ área		63	7,0%	
<b>Habilitações literárias</b> C4	1. 1º ciclo	21	2,3%		4. Coordenador Processo	8	0,9%
	2. 2º ciclo	99	11,1%		5. Gestor/ Técnico	41	4,6%
	3. 3º ciclo	318	35,5%		6. Diretores Loja/ Gestores Negócio	10	1,1%
	4. Ensino secundário	328	36,6%		7. Diretores 1º linha	7	0,8%
	5. Ensino superior	80	8,9%		8. Administrativos, outros	34	3,8%
NS\ NR	39	4,4%	NS\ NR	14	1,6%		

Devido à natureza categórica das variáveis de caracterização, foi necessário proceder à criação de variáveis *dummy*, para mais tarde poderem ser utilizadas no modelo de regressão linear múltipla. Para categoria de referência escolheu-se sempre a categoria modal.

Posto isto, tem-se:

$$\text{Posição hierárquica} = D_{1i} = \begin{cases} 1, & \text{se } C_1 = i, i = 2, \dots, 8 \\ 0, & \text{se } i = 1 \end{cases}$$

$$\text{Classe etária} = D_{2i} = \begin{cases} 1, & \text{se } C_2 = i, i = 1, 2, 4, 5 \\ 0, & \text{se } i = 3 \end{cases}$$

$$\text{Antiguidade na organização} = D_{3i} = \begin{cases} 1, & \text{se } C_3 = i, i = 1, 3, 4, 5 \\ 0, & \text{se } i = 2 \end{cases}$$

$$\text{Habilitações literárias} = D_{4i} = \begin{cases} 1, & \text{se } C_4 = i, i = 1, 2, 3, 5 \\ 0, & \text{se } i = 4 \end{cases}$$

$$\text{Género} = D_{5i} = \begin{cases} 1, & \text{se } C_5 = 1 \\ 0, & \text{se } C_5 = 2 \end{cases}$$

## 5.1.2 Indicadores

Foram realizadas 52 questões na escala de Likert onde 1 representa o nível “Muito baixo” e 10 o nível “Muito alto”. Posteriormente, optou-se por agrupar as questões em indicadores por meio de uma média ponderada de forma a poder ter uma abordagem mais estrutural. Para validar a consistência de cada agrupamento recorreu-se à análise de alfa de Cronbach.

Embora o questionário seja confidencial, a Qmetrics, responsável pela sua elaboração, deu a oportunidade de incluir uma questão por cada indicador criado.

Os indicadores criados foram Infraestrutura Laboral, Envolvimento organizacional, Orientação para o cliente, Relacionamentos interpessoais, Autonomia, Reconhecimento, Desenvolvimento profissional e Remuneração/ benefícios. Na tabela seguinte, encontram-se os indicadores criados juntamente com uma das questões incluídas nesse agrupamento e o resultado do alfa de Cronbach.

*Tabela 3 - Coeficientes alfa de Cronbach*

Indicador	Questões	Alpha de Cronbach
Infraestrutura laboral	- Condições de higiene e segurança das instalações e do local de trabalho - -	0,86
Envolvimento organizacional	- Conhecimento que possui da política e objetivos da qualidade da empresa - -	0,87
Orientação para o cliente	- Grau em que a satisfação dos clientes é uma das maiores prioridades da empresa - -	0,84
Relacionamentos interpessoais	- Ajuda, colaboração e cooperação entre colegas - -	0,87
Autonomia	- Autonomia que é dada para planear, executar e avaliar o próprio trabalho - -	0,88
Reconhecimento	- Sentimento que é feita uma correta e justa avaliação do seu trabalho - -	0,93
Remuneração/ benefícios	- Sentimento que está a ser bem pago - -	0,94
Desenvolvimento profissional	- Oportunidades de progressão profissional na empresa - -	0,79

Todos os alfas de Cronbach são maiores ou igual a 0,79, indicando uma consistência interna robusta.

Podemos observar no gráfico seguinte os valores médios/ desvios padrões da satisfação global e dos indicadores criado.

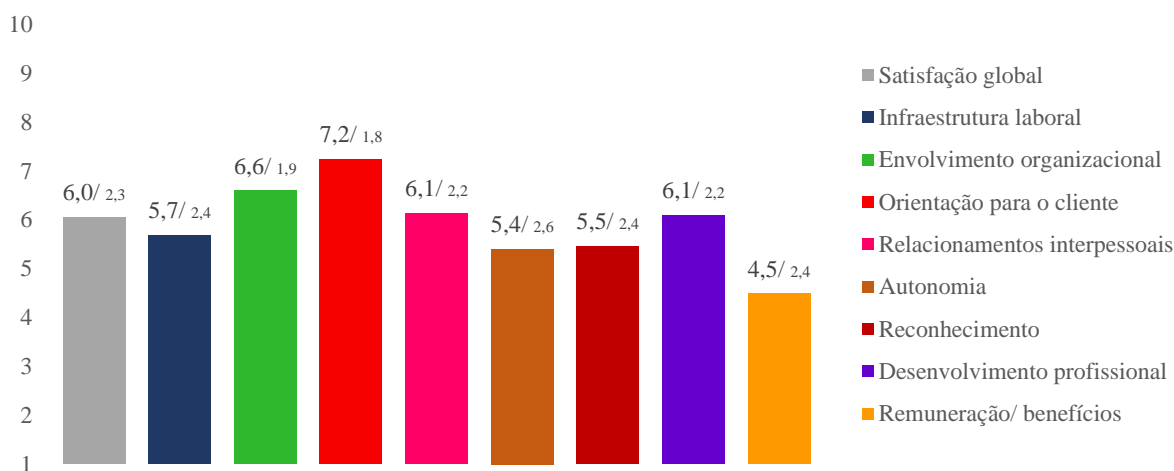


Figura 1 - Valor médio/ desvio padrão por indicador

Por ordem decrescente de pontuação média atribuída aos indicadores tem-se: Orientação para o cliente (7.2), Envolvimento organizacional (6.6), Relacionamentos interpessoais (6.1), Desenvolvimento profissional (6.1), Satisfação global (6.0), Infraestrutura laboral (5.7), Reconhecimento (5.5), Autonomia (5.4) e, por fim, Remuneração/ benefícios (4.5).

Os desvios padrões variam entre 1.8 (Orientação para o cliente) e 2.6 (Autonomia).

## 5.2 Análise comparativa

Este subcapítulo tem como objetivo averiguar a existência de diferenças significativas por estrato em relação aos indicadores. Antes de realizar testes estatísticos é feita uma análise baseada puramente em médias amostrais.

A seguinte tabela mostra a classificação média de todos os indicadores em estudo por cada variável de caracterização.

Tabela 4 - Valor médio por estrato

Variável	Categoria	Satisfação global	Infraestrutura laboral	Envolvimento organizacional	Orientação para o cliente	Relacionamentos interpessoais	Autonomia	Reconhecimento	Desenvolvimento profissional	Remuneração/benefícios	Total
<b>Género</b>	Masculino	6,2	6,1	6,8	7,1	6,5	5,8	5,8	6,0	4,6	6,1
	Feminino	6,0	5,5	6,5	7,3	6,0	5,2	5,3	6,1	4,4	5,8
<b>Classe etária</b>	18-25 anos	6,5	6,4	6,9	7,5	6,9	6,0	6,1	6,5	5,3	6,5
	26-35 anos	6,0	5,8	6,6	7,3	6,2	5,6	5,5	6,0	4,5	5,9
	36-45 anos	5,7	5,2	6,4	7,0	5,6	4,9	5,1	5,8	3,9	5,5
	46-55 anos	6,3	5,7	6,8	7,4	6,2	5,4	5,5	6,4	4,7	6,0
	+55 anos <sup>1</sup>	6,6	5,8	6,5	7,5	6,1	5,5	5,9	6,6	5,4	6,2
<b>Habilitações literárias</b>	1º ciclo <sup>1</sup>	6,2	5,3	6,5	7,3	5,7	4,9	5,2	6,8	4,3	5,8
	2º ciclo	5,6	5,2	6,2	6,8	5,5	4,9	5,0	5,9	4,4	5,5
	3º ciclo	5,8	5,4	6,5	7,1	5,7	5,0	5,2	5,9	4,5	5,7
	Ensino secundário	6,1	5,8	6,6	7,4	6,4	5,6	5,6	6,1	4,6	6,0
	Ensino superior	7,2	6,8	7,6	7,6	7,7	7,1	6,8	6,6	4,5	6,9
<b>Antiguidade na organização</b>	< 1 ano	7,0	7,0	7,2	7,6	7,4	6,3	6,7	7,1	5,9	6,9
	1-5 anos	6,0	5,7	6,7	7,2	6,2	5,7	5,4	5,8	4,5	5,9
	6-10 anos	5,7	5,8	6,5	7,2	6,4	5,5	5,4	6,0	3,9	5,8
	11-20 anos	5,4	4,9	6,1	7,0	5,3	4,7	4,8	5,6	3,7	5,3
	> 20 anos	6,0	5,3	6,5	7,2	5,8	5,1	5,2	6,2	4,3	5,7
<b>Posição hierárquica</b>	Operador/ Vendedor	5,8	5,4	6,3	7,1	5,8	4,9	5,1	5,9	4,4	5,6
	Supervisor	7,3	6,8	7,3	7,9	6,6	6,5	6,0	7,2	5,2	6,8
	Coordenador equipa/ área	7,1	6,7	7,7	8,1	7,2	7,3	7,1	7,5	5,5	7,1
	Coordenador Processo <sup>2</sup>	6,1	8,1	7,3	6,9	7,5	6,4	6,8	6,7	5,3	6,8
	Gestor/ Técnico	7,1	7,2	7,7	7,9	7,7	7,4	6,5	6,2	3,9	6,8
	Diretores Loja/ Gestores Negócio <sup>1</sup>	8,4	6,7	8,2	8,4	8,1	8,5	8,3	7,9	6,4	7,9
	Diretores 1º linha <sup>2</sup>	8,9	7,8	8,9	8,4	8,7	9,1	8,7	8,0	7,3	8,4
Administrativos, Outros	6,3	5,8	7,0	7,4	6,5	6,1	5,8	6,3	3,8	6,1	

**Género** - Observa-se que os homens atribuem classificações mais altas em todos os indicadores com exceção da orientação para o cliente e desenvolvimento profissional. É na infraestrutura laboral e nos relacionamentos interpessoais onde essa diferença é maior (0.6 e 0.5, respetivamente).

**Classe etária** - A classe 36-45 anos regista os menores valores médios em todos os indicadores. No lado oposto, a classe 18-25 anos tem os maiores valores médios na maioria das categorias.

**Habilitações literárias** - Observa-se, excluindo a classe 1ª ciclo por ter uma pequena dimensão (inferior a 30), que quanto maior o grau académico, maior são os valores dados em todos os indicadores.

**Antiguidade na organização** - Os trabalhadores que estão na organização há 11-20 anos são os que dão classificações mais baixas nos indicadores estudados. No lado oposto, os que estão há menos de um ano são os que dão classificações mais altas.

**Posição hierárquica** - Como já foi visto anteriormente, a grande maioria da amostra pertence à classe Operador/ Vendedor, ainda assim, existem outras categorias com uma dimensão superior a 30 (valor muitas vezes usado na literatura para delimitar a amostra de pequena dimensão da amostra de grande dimensão). Dentro das categorias com uma dimensão superior a 30, a categoria Operador/ Vendedor é a que dá as menores classificações em praticamente todos indicadores.

## 5.2.1 Teste Kruskal-Wallis – Post Hoc de Dunn

Utilizou-se o teste de Kruskal-Wallis<sup>1</sup> para averiguar a existência de diferenças entre grupos em relação à satisfação global e aos indicadores. Aquando da existência de diferenças significativas, realizou-se o teste *post hoc*, o teste de Dunn, de forma a poder identificar quais os grupos responsáveis pelas diferenças encontradas.

<sup>1</sup> Dimensão inferior a 30.

<sup>2</sup> Dimensão inferior a 10.





### 5.3 Aplicação do Modelo de Regressão Linear múltipla

Serve o presente capítulo para encontrar um modelo explicativo para a satisfação global. Para tal, foi utilizado o *software* SPSS.

Para a seleção de variáveis foi aplicado o método *stepwise* com critério de entrada valor  $p < 0,05$  e critério de saída valor  $p > 0,10$  (na estatística F). Nesta análise foram excluídas todas as observações que tenham pelo menos um valor *missing* em uma das suas variáveis.

As variáveis predictoras são Infraestrutura laboral, Envolvimento organizacional, Orientação para o cliente, Relacionamentos interpessoais, Autonomia, Reconhecimento, Desenvolvimentos profissional, Remuneração/ benefícios e as variáveis *dummy* referentes à Classe etária, Antiguidade na organização, Hbilitações literárias, Posição hierárquica e Género.

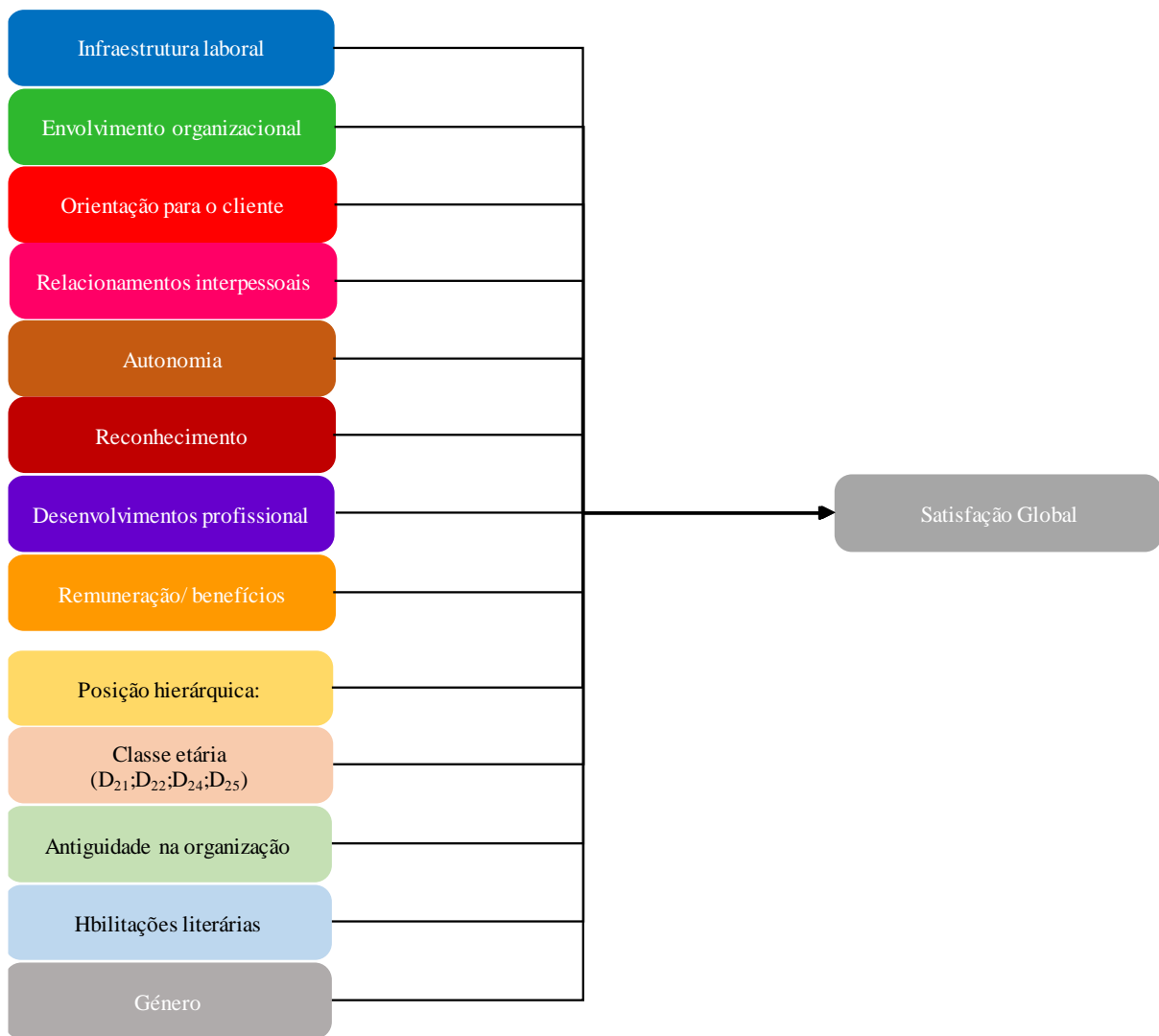


Figura 8 – Representação esquemática do modelo de regressão linear múltipla

Aplicado o algoritmo *stepwise*, obteve-se um modelo que revelou ter pontos influentes, identificados através da distância de Cook.

Tabela 5 - Pontos influentes do primeiro modelo

Número do caso	Erro Resíduo	SGlob	Valor previsto	Resíduo
94	3,27	10	5,40	4,60
98	4,01	10	4,37	5,63
110	3,54	10	5,02	4,98
204	3,30	10	5,36	4,64
337	-3,56	2	7,01	-5,01
400	5,68	10	2,02	7,98
401	4,55	10	3,60	6,40
413	-3,36	3	7,72	-4,72
567	3,57	9	3,98	5,02

Analisou-se a normalidade dos resíduos do através do teste Kolmogorov-Smirnov e obteve-se um valor p inferior a 0.001, rejeitando a hipótese da normalidade dos resíduos.

Tabela 6 - Teste Kolmogorov-Smirnov aos resíduos do primeiro modelo

	Kolmogorov-Smirnov		
	Estatística	gl	Sig.
Standardized Residual	0,046	873	<,001

Posto isto, decidiu-se remover os pontos influentes e aplicar novamente o algoritmo.

Aplicado o algoritmo *stepwise*, ele parou na 7ª iteração.

Embora o algoritmo permitisse a retirada de variáveis, ele nunca o fez. As variáveis adicionadas foram, por ordem de entrada, Envolvimento organizacional (EOrg), Desenvolvimento profissional (DPro), Relacionamentos interpessoais (RInter), Remuneração/ benefícios (Rben), Infraestrutura laboral (ILab), D<sub>41</sub> e Reconhecimento (Rec). As variáveis não incluídas no modelo foram Orientação para o cliente, (OClie) Autonomia (Aut), e as restantes variáveis *dummy*.

Uma das conclusões a tirar é que o facto do Envolvimento organizacional ser a primeira variável a ser adicionada significa que, isoladamente, é a variável Envolvimento organizacional que melhor explica a satisfação global.

À medida que vão sendo adicionadas variáveis ao modelo, a estimativa do erro padrão vai diminuindo e tanto o  $R^2$  como o  $R^2$  ajustado vão incrementando, como seria de esperar. No final obtemos um modelo com um  $R^2$  ajustado de 0,688 e uma estimativa de erro padrão de 1,297.

Tabela 7 – Estatísticas de mudança dos modelos criados

Modelo	R <sup>2</sup>	R <sup>2</sup> ajustado	Erro padrão da estimativa	Estatísticas de mudança			
				Variável adicionada	Mudança de R <sup>2</sup>	Mudança F	Sig. Mudança F
1	0,572	0,571	1,521	Envolvimento organizacional	0,572	1004,604	0,000
2	0,641	0,640	1,393	Desenvolvimento profissional	0,070	145,773	0,000
3	0,672	0,670	1,334	Relacionamentos interpessoais	0,030	69,424	0,000
4	0,682	0,680	1,314	Remuneração/ benefícios	0,010	23,730	0,000
5	0,687	0,685	1,303	Infraestrutura laboral	0,006	13,638	0,000
6	0,689	0,687	1,300	D41	0,002	4,894	0,027
7	0,691	0,688	1,297	Reconhecimentos	0,002	4,014	0,045

A análise de variância diz-nos que o modelo é significativo, isto é, para o nível de significância estudado podemos admitir que pelo menos uma das variáveis do modelo tem influência na satisfação global.

Tabela 8 - Tabela Anova do modelo final

Modelo	Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
9 Regressão	2809,54	7	401,36	238,61	<,001
Resíduo	1256,55	747	1,68		
Total	4066,09	754			

Na tabela seguinte, encontram-se os coeficientes do modelo, juntamente com algumas estatísticas associadas ao modelo. Nela observa-se que os coeficientes associados às variáveis são significativos, o que seria de esperar dado o critério utilizado no algoritmo. Mantendo as restantes variáveis constantes, observa-se que um aumento de uma unidade no Desenvolvimento profissional resultará em um aumento de 0.29 na Satisfação global. Seguindo o mesmo raciocínio, a Satisfação global é incrementada em 0.22 para um aumento de uma unidade no Envolvimento organizacional, 0.18 para os Relacionamentos interpessoais, 0.11 para Remuneração/ benefícios, 0.10 para a infraestrutura laboral e 0.09 para o Reconhecimento.

Em relação à interpretabilidade das variáveis *dummy*, podemos afirmar que mantendo as restantes variáveis constantes, o grau de satisfação dos colaboradores com o Ensino secundário é superior em 0.71 aos colaboradores com apenas o 1º ciclo (para uma significância de 95%).

O modelo não tem problemas de multicolinearidade já que os VIF's associados aos coeficientes são menores que 10.

Tabela 9 - Resumo do modelo final

Modelo	Coeficientes não padronizados		Coeficientes padronizados Beta	t	Sig.	95,0% Intervalo de Confiança para B		Estatísticas de colinearidade	
	B	Erro Erro				Limite inferior	Limite	Tolerância	VIF
9 (Constante)	0,17	0,17		0,99	0,32	-0,16	0,50		
Envolvimento organizacional	0,22	0,05	0,19	4,70	<,001	0,13	0,31	0,26	3,88
Desenvolvimento profissional	0,29	0,04	0,27	7,51	<,001	0,22	0,37	0,32	3,15
Relacionamentos interpessoais	0,18	0,04	0,18	4,29	<,001	0,10	0,27	0,24	4,14
Remuneração/ benefícios	0,11	0,03	0,11	3,51	<,001	0,05	0,17	0,41	2,44
Infraestrutura laboral	0,10	0,03	0,11	3,34	<,001	0,04	0,16	0,42	2,37
D41	-0,71	0,32	-0,05	-2,20	0,03	-1,34	-0,08	0,98	1,02
Reconhecimento	0,09	0,04	0,09	2,00	0,05	0,00	0,18	0,21	4,84

A satisfação global pode então ser explicada pela seguinte equação:

$$SGlob = 0.17 + 0.22 * EOrg + 0.29 * DPro + 0.18 * RInt + 0.11 * RBen + 0.10 * ILab - 0.71 * D41 + 0.09 * Rec$$

Ainda que estejamos perante um modelo com coeficientes estatisticamente significativos, ausente de multicolinearidade e com um  $R^2$  de 0.69, ele pode não atender aos pressupostos, interferindo com a qualidade do mesmo. No seguinte subcapítulo avalia-se a qualidade do modelo de forma a garantir que se trata de um modelo fiável.

### 5.3.1 Avaliação do modelo

A análise dos resíduos é uma peça fundamental na validação do modelo. Este subcapítulo tem como objetivo validar a normalidade e autocorrelação dos resíduos. Uma vez provada a normalidade e autocorrelação dos resíduos, as inferências do modelo tornam-se mais confiáveis.

#### Normalidade dos resíduos

No histograma seguinte, observa-se que os resíduos apresentam uma distribuição simétrica em torno de zero e quanto maior a distância a zero, menor a frequência dos mesmos.

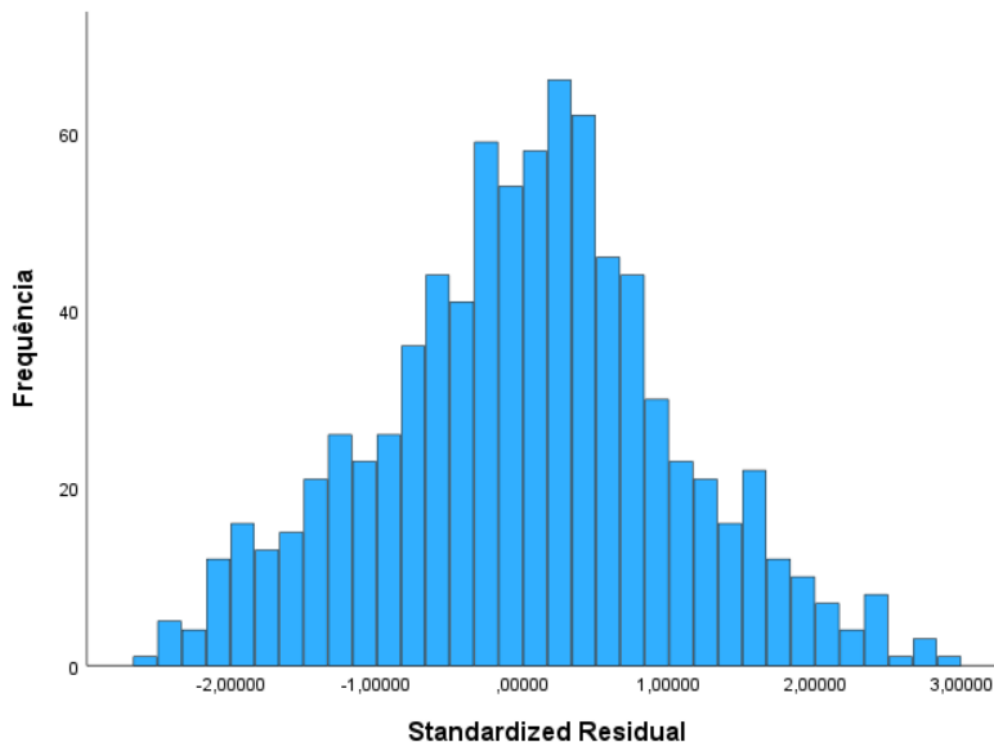


Figura 9 - Histograma resíduos do modelo final

Apesar da análise gráfica ser favorável, realizou-se o teste de Kolmogorov-Smirnov de forma a ter evidências quantitativas. O valor p resultante foi de 0.12, ou seja, não se rejeita a normalidade dos resíduos.

Tabela 10 - Teste Kolmogorov-Smirnov aos resíduos do modelo final

	Kolmogorov-Smirnov		
	Estatística	gl	Sig.
Standardized Residual	0,028	830	0,120

### Autocorrelação dos resíduos

A presença de autocorrelação nos resíduos contraria os pressupostos do modelo linear. Assim sendo, é essencial analisar a autocorrelação dos resíduos.

Para analisar a autocorrelação foi calculada a estatística de Durbin-Watson, obtendo o valor de 1.91. A estatística de Durbin-Watson varia entre 0 e 4, onde valores próximos de 2 sugerem a ausência de autocorrelação. Assim, com um valor próximo a 2, podemos concluir que os resíduos não apresentam autocorrelação.

Tabela 11 - Teste Durbin-Watson aos resíduos

	Estatística
Durbin-Watson	1,928

## Conclusão

O principal objetivo deste projeto foi desenvolver um modelo de regressão linear múltipla explicativo para a satisfação global laboral.

O modelo obtido apresenta um bom ajustamento ( $R^2 \approx 0.69$ ) e está em conformidade com os usuais pressupostos da regressão linear, permitindo tirar conclusões fiáveis do mesmo. O que o modelo indica é que quanto maior o envolvimento do colaborador com a sua organização, maior a sua satisfação laboral. De forma análoga, a satisfação laboral do colaborador também irá aumentar quando aumentar: (1) o sentimento de que se estão a desenvolver profissionalmente; (2) a satisfação com os seus relacionamentos no local de trabalho; (3) a satisfação com a sua remuneração/ benefícios recebidos; (4) a satisfação com as infraestruturas/ equipamentos e (5) a perceção de que o seu trabalho está a ser reconhecido.

Fora do modelo ficaram as variáveis Orientação para o cliente, Autonomia, Posição hierárquica, Classe etária, Antiguidade na organização e Género. Ou seja, a forma como o colaborador percebe os clientes e o seu grau de autonomia não tiveram impactos significativos na sua satisfação laboral. Além destes indicadores, as variáveis de caracterização Posição hierárquica, Classe etária, Antiguidade na organização e Género também não revelaram ter impacto na satisfação laboral.

Foi realizada uma análise comparativa dos indicadores por estrato através de testes estatísticos. Enumerar todas as diferenças significativas encontradas seria bastante exaustivo por haver um número elevado de indicadores e variáveis de caracterização. Assim sendo, é reportado apenas se duas categorias registam diferenças significativas em pelo menos um indicador, indicando o que atribui maiores classificações.

Foram identificadas diferenças significativas por género, sendo o género masculino aquele que atribui classificações mais elevadas. Os colaboradores com o grau de escolaridade maior ou igual ao ensino superior atribuem classificações mais altas do que aqueles com o grau de ensino secundário, 3º ciclo e 2º ciclo. Aqueles que se encontram na empresa há menos de um ano registam classificações mais elevadas que os restantes, os que estão na empresa há 1-5 anos também atribuem classificações mais altas em relação a aqueles que se encontram na empresa há 11-20 anos. Relativamente à idade do colaborador, observam-se diferenças apenas entre os grupos etários 18-25 anos e 35-45 anos, com o primeiro grupo a atribuir classificações mais elevadas. Finalmente, os Operadores/ Vendedores atribuem classificações mais baixas do que Coordenadores de equipa/ área, Gestores/ Técnicos e Diretores de loja/ Gestores de negócio.

Em relação às limitações do estudo, é importante destacar que os dados pertencem exclusivamente a uma única empresa. Considerando que os dados pertencem a uma empresa do setor indústria/ retalho, é plausível afirmar que as conclusões deste estudo possam ser mais aplicáveis a colaboradores dentro desse setor comparativamente a outros.

## Referências Bibliográficas

- [1] ScienceDirect. Job Satisfaction. URL: <https://www.sciencedirect.com/topics/social-sciences/job-satisfaction>.
- [2] Locke. The Nature and Causes of Job Satisfaction
- [3] Peter Warr. Work and Well-Being
- [4] Danica Bakotić. Relationship between Job Satisfaction and Organizational Performance
- [5] Teresa Alpuim. Regressão Linear Simples (Apontamentos da cadeira Modelos Lineares). Faculdade de Ciências da Universidade de Lisboa
- [6] Teresa Alpuim. Regressão Linear Múltipla (Apontamentos da cadeira Modelos Lineares). Faculdade de Ciências da Universidade de Lisboa.
- [7] Douglas C. Montgomery Elizabeth A. Peck G. Geoffrey Vining. Introduction to Linear Regression Analysis.
- [8] Alvin C. Rencher and G. Bruce Schaalje. Linear Models in Statistics
- [9] James K. Lindsey. Applying Generalized Linear Models
- [10] Le Moyne College. Nonparametric Tests. URL: <https://web.lemoyne.edu/courseinformation/mth%20112/rinaman/instman/CH14.PDF>
- [11] Yudhistir S.M.F. Jugessur. Reliability and Internal Consistency of data: Significance of Calculating Cronbach's Alpha Coefficient in Educational Research. URL: [https://www.ijhssi.org/papers/vol11\(4\)/Ser-3/C1104030914.pdf](https://www.ijhssi.org/papers/vol11(4)/Ser-3/C1104030914.pdf)
- [12] IBM SPSS Statistics 28 Brief Guide. [https://www.ibm.com/docs/en/SSLVMB\\_28.0.0/pdf/IBM\\_SPSS\\_Statistics\\_Brief\\_Guide.pdf](https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Brief_Guide.pdf)