Analytics e Inteligência Artificial Data Science

Tema da aula
Regressão Linear

FIA BUSINESS SCHOOL lab.data

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

BUSINESS SCHOOL lab.data

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil.

Os diretores foram professores de grandes especialistas do mercado.

+10 anos de atuação.

+9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases;*
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- **ANALYTICS**
- INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- ANÁLISE EXPLORATÓRIA DE DADOS
- INFERÊNCIA ESTATÍSTICA
- **TÉCNICAS DE PROJEÇÃO**
- TÉCNICAS DE CLASSIFICAÇÃO
- TÓPICOS DE MODELAGEM
- TÉCNICAS DE SEGMENTAÇÃO
- TÓPICOS DE ANALYTICS
- MANIPULAÇÃO DE BASE DE DADOS
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- **LINGUAGEM R**
- LINGUAGEM PYTHON
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

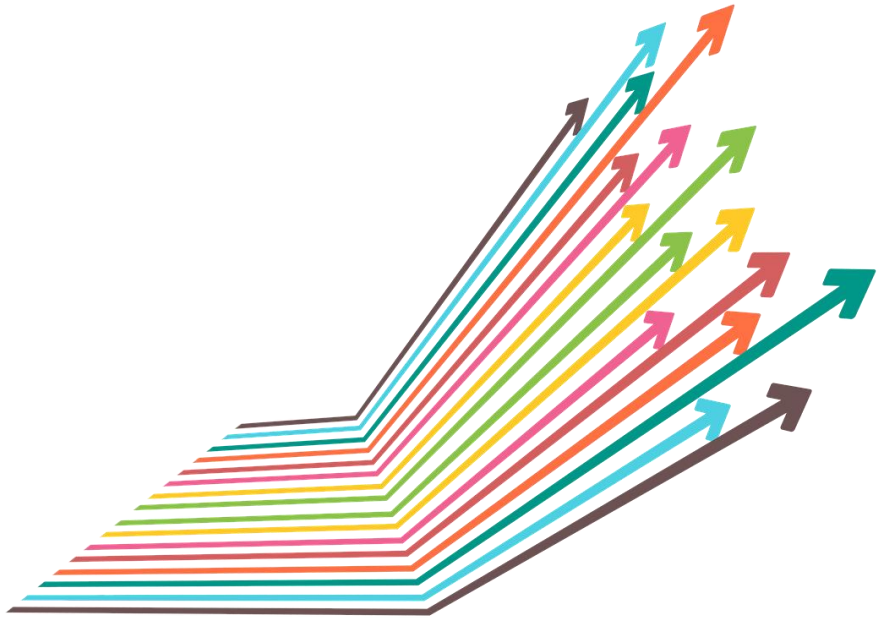
1. Introdução: Modelagem Supervisionada
2. Objetivo
3. Coeficiente de Correlação Linear
4. Equação da Reta
   - Método de Mínimos Quadrados
   - Resíduos
5. Regressão Linear Simples
   - Intervalo de Confiança para β_0 e β_1
   - Abordagem por Teste de Hipóteses
   - Diagnóstico do Modelo
6. Regressão Linear Múltipla
   - Análise Bidimensional: Correlograma
   - Qualidade de Ajuste
   - Processo de Seleção de Variáveis
   - Colinearidade
   - Incorporando Variáveis Qualitativas

Referências Bibliográficas

1. Introdução: Modelagem Supervisionada

O Que é um Modelo Estatístico?

- ✓ Um modelo estatístico é uma ferramenta que ajuda a explicar as **relações existentes** entre duas ou mais variáveis em uma **população**, a partir de dados de uma **amostra**.
- ✓ Frequentemente, essa tarefa é realizada por meio de uma **equação matemática** ou um **algoritmo** que descreve como uma determinada variável de interesse *(resposta)* se comporta a depender de outras variáveis relacionadas a ela *(explicativas)*.
- ✓ Com base nessa equação ou algoritmo, é possível, também, **realizar previsões** sobre como possíveis mudanças de valores nas variáveis explicativas afetarão a variável resposta para novas observações.
- ✓ Esse é o tipo de modelo que chamamos de **supervisionado**. Quanto melhor for a predição da variável resposta a partir das variáveis explicativas, melhor será o modelo.



FIA BUSINESS SCHOOL lab.data

O Que é um Modelo Estatístico?

Um modelo estatístico é como uma receita culinária, que descreve a maneira pela qual um **resultado final** *(variável resposta)* pode ser obtido a partir dos **ingredientes** *(variáveis explicativas)*.

Bolo de cenoura com chocolate

3 cenouras médias

4 ovos

2 ½ xícaras de farinha de trigo

1 colher de sopa de fermento em pó

Entre outros...

O Que é um Modelo Estatístico?

Um modelo estatístico é como uma receita culinária, que descreve a maneira pela qual um **resultado final** *(variável resposta)* pode ser obtido a partir dos **ingredientes** *(variáveis explicativas)*.

O que acontece se mudarmos as medidas associadas aos ingredientes?

Bolo de cenoura com chocolate

7 cenouras médias

2 ovos

½ xícara de farinha de trigo

5 colheres de sopa de fermento em pó

Entre outros...

O Que é um Modelo Estatístico?

Um modelo estatístico é como uma receita culinária, que descreve a maneira pela qual um **resultado final** *(variável resposta)* pode ser obtido a partir dos **ingredientes** *(variáveis explicativas)*.

E se retirarmos alguns ingredientes?

Bolo de cenoura com chocolate

3 cenouras médias

4 ovos

2 ½ xícaras de farinha de trigo

1 colher de sopa de fermento em pó

Entre outros...

O Que é um Modelo Estatístico?

Um modelo estatístico é como uma receita culinária, que descreve a maneira pela qual um **resultado final** *(variável resposta)* pode ser obtido a partir dos **ingredientes** *(variáveis explicativas)*.

E se modificarmos alguns ingredientes?

Bolo de cenoura com chocolate

3 cenouras médias

4 ovos

4 fatias de queijo 🤔

2 ½ xícaras de farinha de trigo

1 colher de sopa de fermento em pó

Entre outros...

FIA BUSINESS SCHOOL lab.data

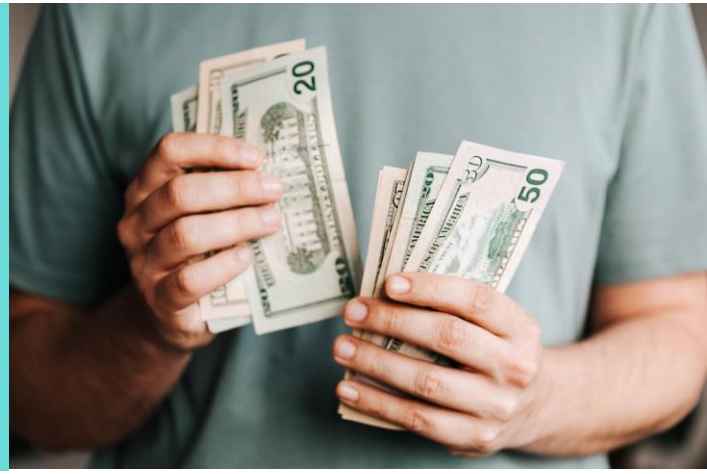
O Que é um Modelo Estatístico?

- ✓ Fica evidente que não é qualquer combinação de ingredientes que resultará num bolo de cenoura com chocolate. Algumas variações podem dar certo, outras não. Para obter o resultado final esperado, é necessário seguir ingredientes apropriados, em proporções apropriadas.
- ✓ De forma análoga, precisamos identificar quais são as **variáveis explicativas** apropriadas, com seus respectivos **pesos** (medidas de importância), para obter a predição correta de uma variável resposta de interesse.

Case: Limite de Cheque Especial

Exemplo:
Determinar o valor de limite de cheque especial ideal para cada cliente correntista, em função de sua renda mensal, padrões de transacionalidade na conta corrente e no cartão de crédito, tempo de relacionamento etc., a fim de gerar maior rentabilização e diminuir a inadimplência.

Aplicação:
Área de crédito em bancos



FIA BUSINESS SCHOOL lab.data

Case: Desempenho Escolar

Exemplo:
Entender a relação entre a nota individual de alunos de escolas públicas no ENEM e características das escolas onde cada um estudou no Ensino Médio, tais como região, presença de recursos tecnológicos, nível de qualificação dos professores etc., a fim de otimizar as políticas de investimento educacional.

Aplicação:
Área de educação pública

Case: Colesterol

Exemplo:
Analisar a relação entre a pressão arterial de pacientes que tomam um medicamento anti-hipertensivo e fatores como ingestão de sal, nível de potássio no sangue, prática de atividade física, IMC etc., a fim de estabelecer medidas de conscientização a respeito da saúde desse grupo.

Aplicação:
Área de saúde

FIA BUSINESS SCHOOL lab.data

Case: Vendas em Varejo

Exemplo:
Entender como determinados fatores estão associados ao volume de vendas em cada filial de um varejista, tais como preço médio dos produtos, variedade de itens disponíveis, localização da filial, nível de satisfação dos clientes etc., buscando atuar na solução de problemas que exerçam impacto negativo sobre as vendas.

Aplicação:
Área comercial, área de varejo

Case: Faturamento em *E-commerce*

Exemplo:
Avaliar se o volume (R$) de investimento em anúncios de mídias digitais para cada produto de um e-commerce está associado ao seu faturamento bruto em determinado período, a fim de otimizar estratégias de *marketing*.

Aplicação:
Área de *e-commerce*

FIA BUSINESS SCHOOL lab.data

Case: Tempo de Internação

Exemplo:
Predizer o tempo de internação de cada paciente em internação hospitalar, com base em características diversas do seu perfil e da sua patologia, a fim de estabelecer um gerenciamento mais eficaz da operação e alocar recursos de forma apropriada.

Aplicação:
Área de gestão hospitalar

2. Objetivo

Objetivo

O objetivo do **modelo de regressão linear** consiste em **explicar** ou **predizer** o valor de uma característica quantitativa a depender de uma ou mais variáveis potencialmente associadas a ela de forma **linear**.

Para isso, teremos que lançar mão do conceito matemático de **função linear** (ou **função de primeiro grau**) para um conjunto de duas ou mais variáveis.

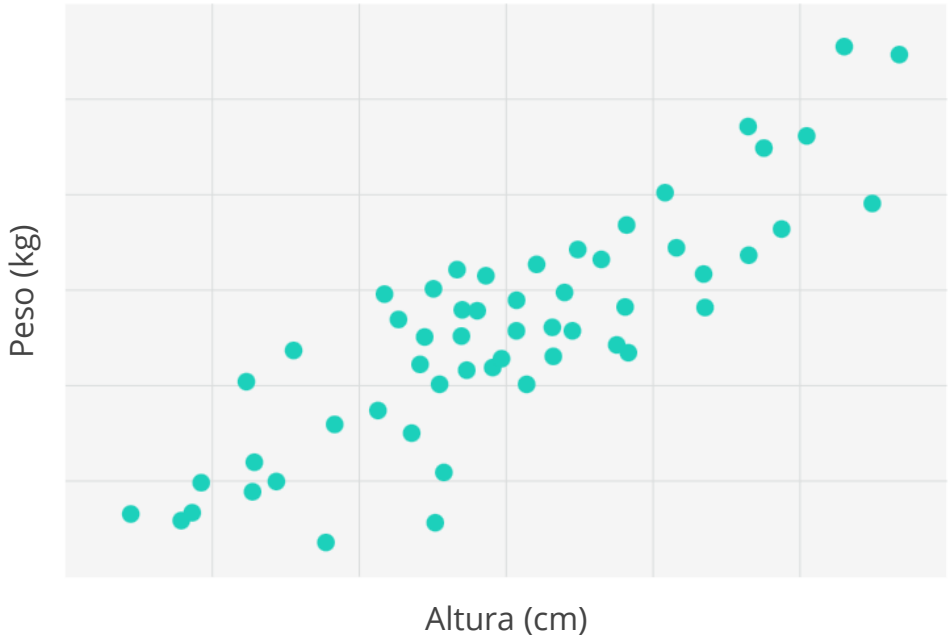
Nesta aula, vamos estudar os aspectos teóricos acerca da regressão linear, bem como entender como aplicá-la em diferentes *cases* práticos.

3. Coeficiente de Correlação Linear

O Que é uma Relação Linear?

Dizemos que duas variáveis quantitativas possuem **relação linear** (ou ***correlação linear***) se, quando o valor de uma das variáveis aumenta, o valor da outra variável também aumenta em uma **proporção constante**.

Isso pode ser ilustrado em um gráfico de dispersão, onde o comportamento dos pontos se assemelha ao de uma **linha reta**.



*Exemplo clássico de **relação linear** entre peso e altura de indivíduos independentes.*

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

Tempo de experiência (em anos)	Valor médio mensal das vendas (em milhares de R$)
1	775
1	630
2	775
2	1.046
3	752
3	255
3	1.049
3	701
3	418
4	871
5	1.340
5	730
5	1.578
6	580
7	967

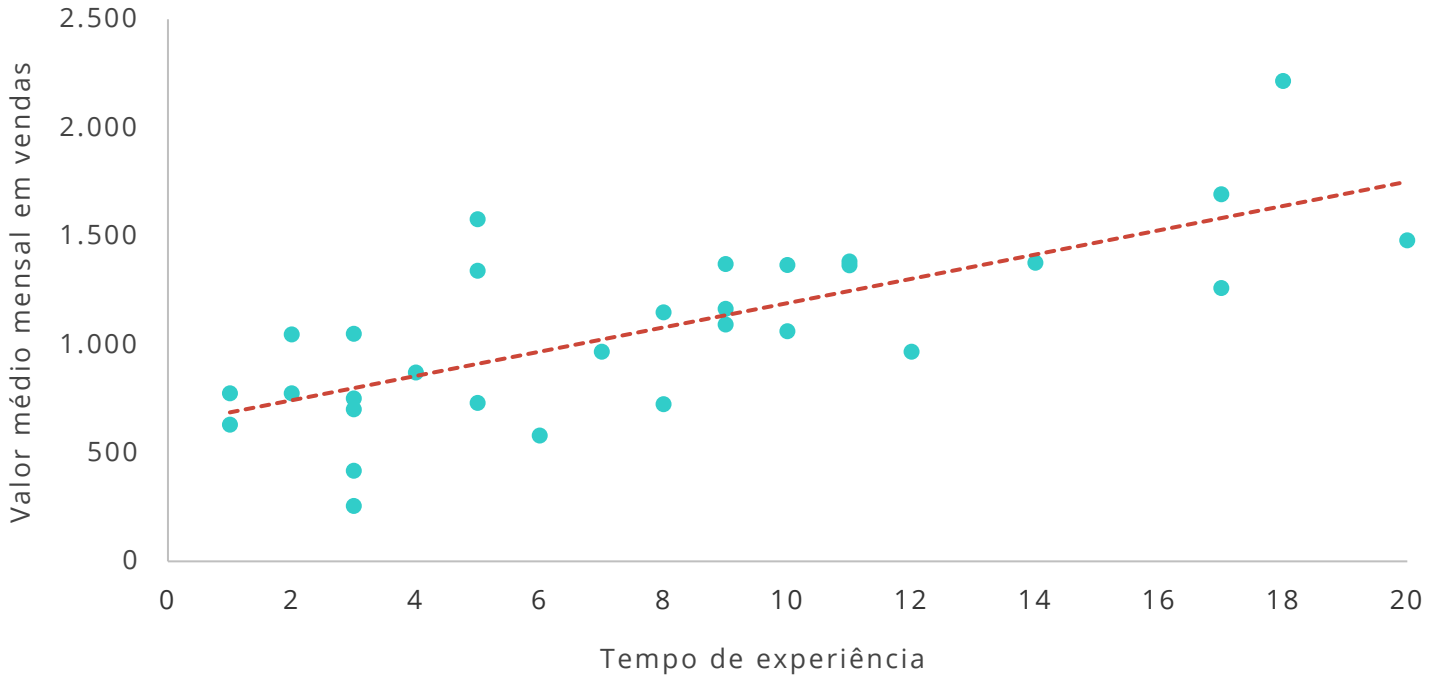
Tempo de experiência (em anos)	Valor médio mensal das vendas (em milhares de R$)
8	1.148
8	724
9	1.371
9	1.165
9	1.092
10	1.061
10	1.367
11	1.365
11	1.383
12	967
14	1.378
17	1.693
17	1.260
18	2.215
20	1.481

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

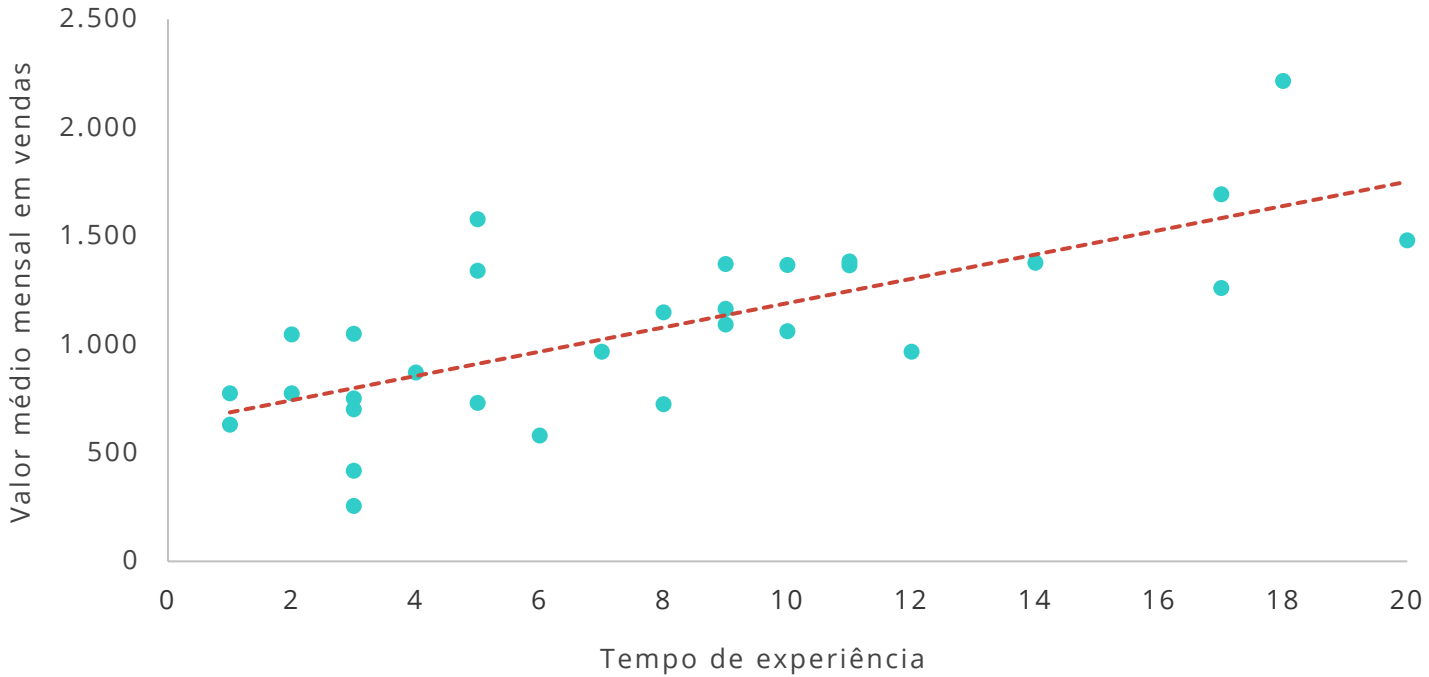
Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



A relação entre o tempo de experiência e o valor das vendas pode ser aproximado de forma razoável a partir de uma reta.

Será que existe alguma forma de mensurar a "força" dessa relação?

Arquivos: Venda_Veiculos (.xlsx e .txt)

Coeficiente de Correlação Linear

O **coeficiente de correlação linear** (denotado pela letra r) mensura o grau de relação linear entre duas variáveis quantitativas *x* e *y*.

Este coeficiente varia entre -1 e 1, sendo que:

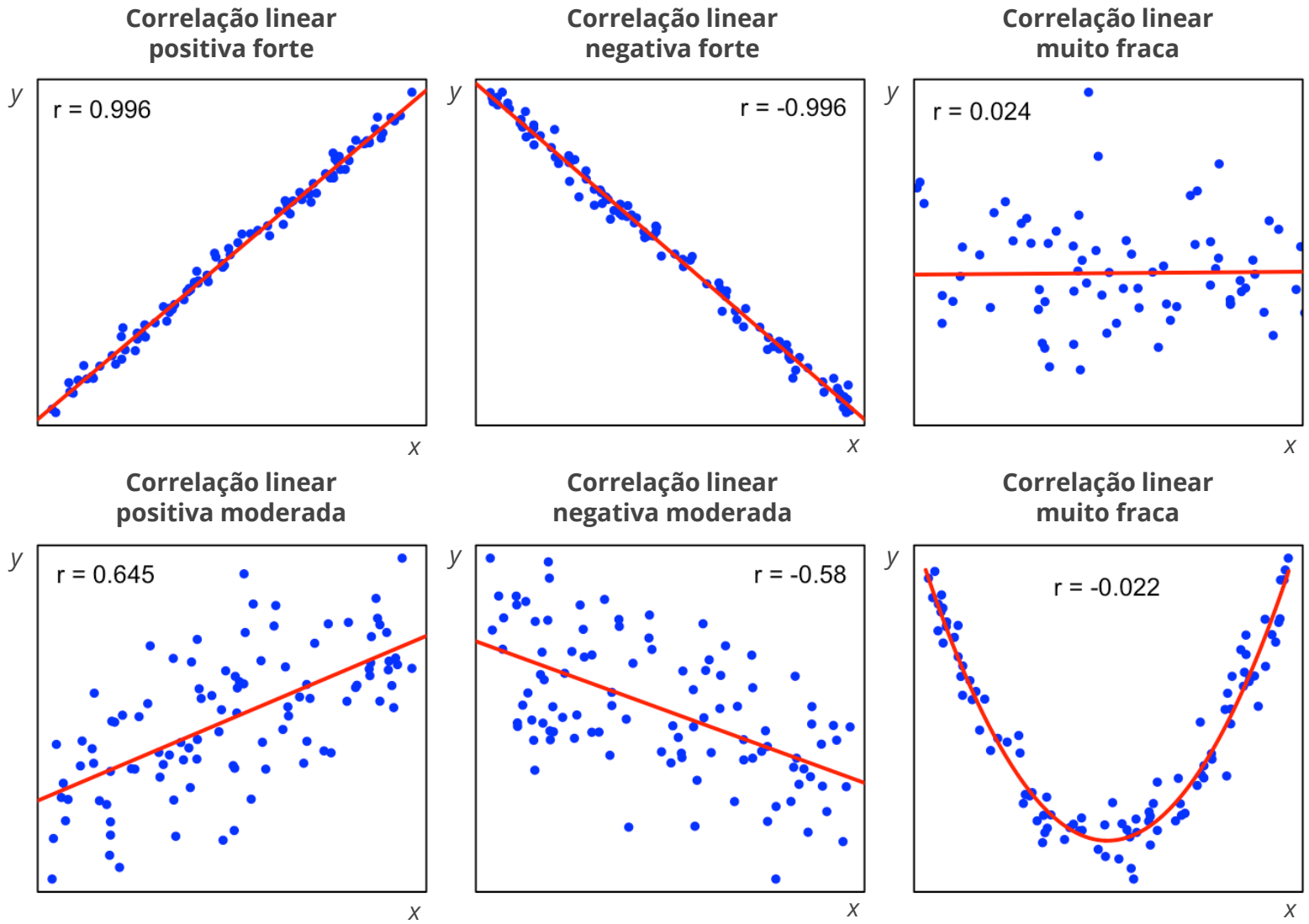
- Quanto mais próximo de **1**, maior a correlação linear **positiva** (ou diretamente proporcional) entre *x* e *y*.
- Quanto mais próximo de **-1**, maior a correlação linear **negativa** (ou inversamente proporcional) entre *x* e *y*.
- Quanto mais próximo de **0**, menor a correlação linear entre *x* e *y*.

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

Fórmula do coeficiente de correlação linear

Karl Pearson (Londres, 1857-1936), estatístico que criou o conceito do coeficiente de correlação linear e fundou o primeiro departamento universitário de Estatística da história.

Interpretação dos Valores de Correlação



Sugestão de interpretação

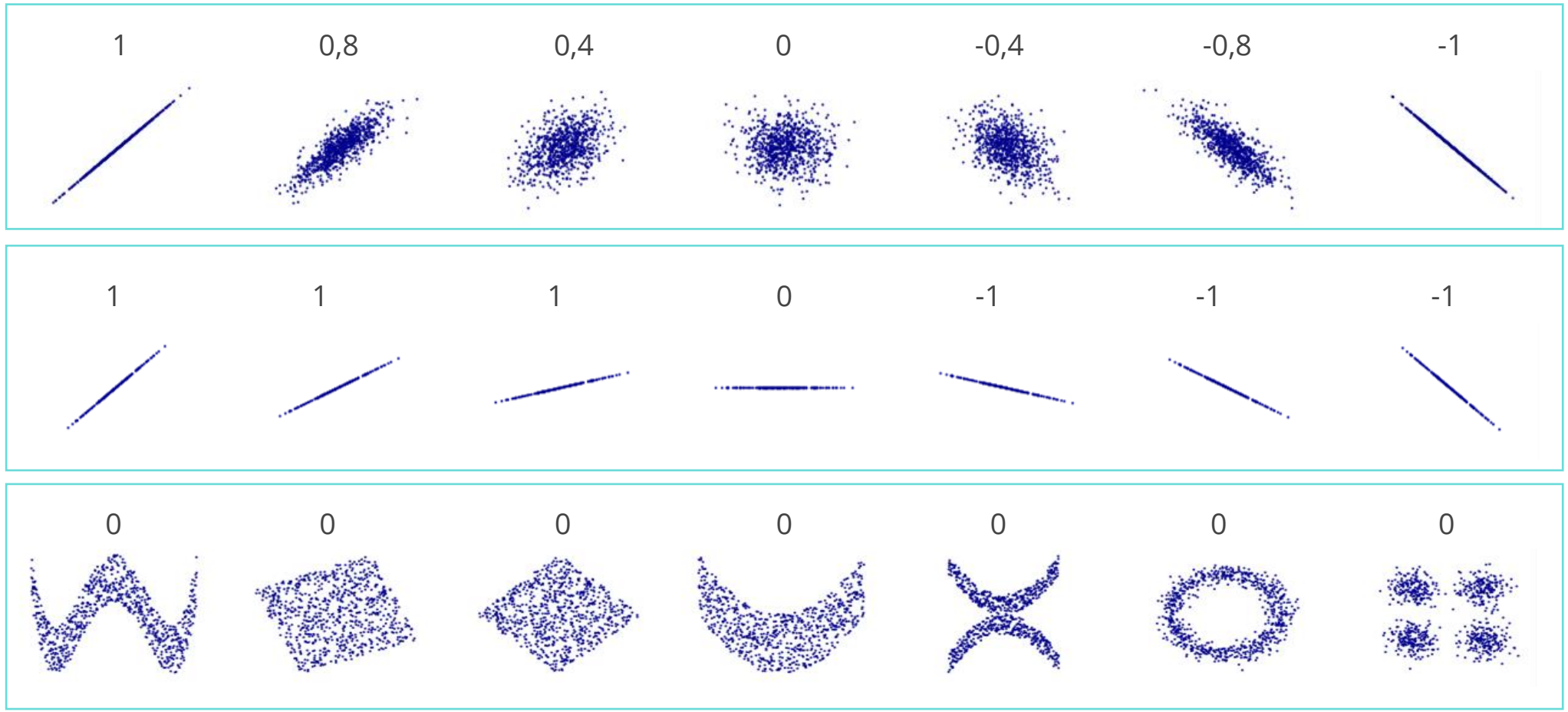
Valor	Relação linear
$\|r\| \geq 0{,}7$	Forte
$0{,}4 \leq \|r\| < 0{,}7$	Moderada
$0{,}2 \leq \|r\| < 0{,}4$	Fraca
$\|r\| < 0{,}2$	Muito fraca

Nota: a interpretação da **força** da correlação linear é subjetiva e pode variar a depender do contexto.

Fonte da imagem: Department of Earth Sciences - Freie Universität Berlin

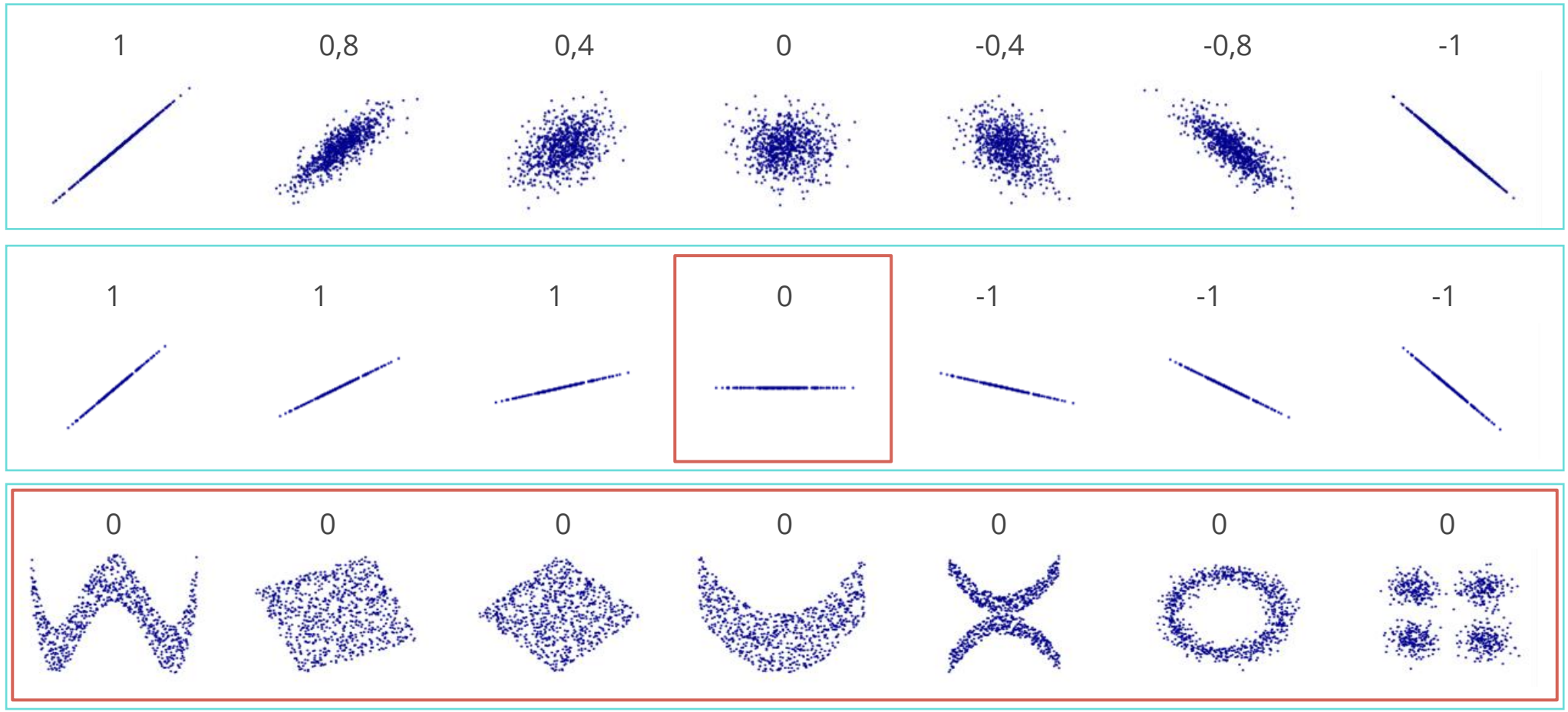
FIA BUSINESS SCHOOL lab.data

Interpretação dos Valores de Correlação



Créditos da imagem: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

Interpretação dos Valores de Correlação



Reta com inclinação horizontal indica ausência de correlação linear!

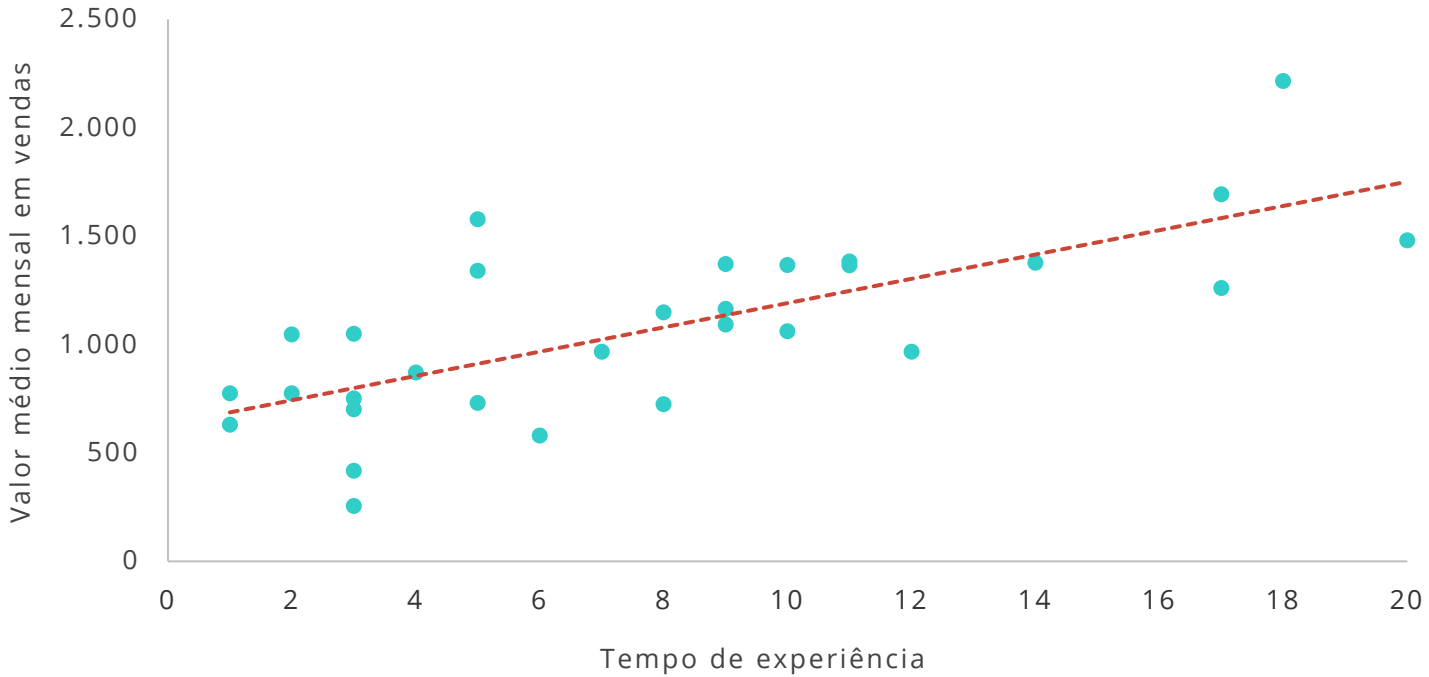
Coeficiente r = 0 não impede que relações não lineares possam existir!

Créditos da imagem: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



Existe uma forte correlação (r **= 0,726**) entre as duas variáveis, ou seja, quanto maior o tempo de experiência do vendedor, maior tende a ser o seu valor médio mensal em vendas.

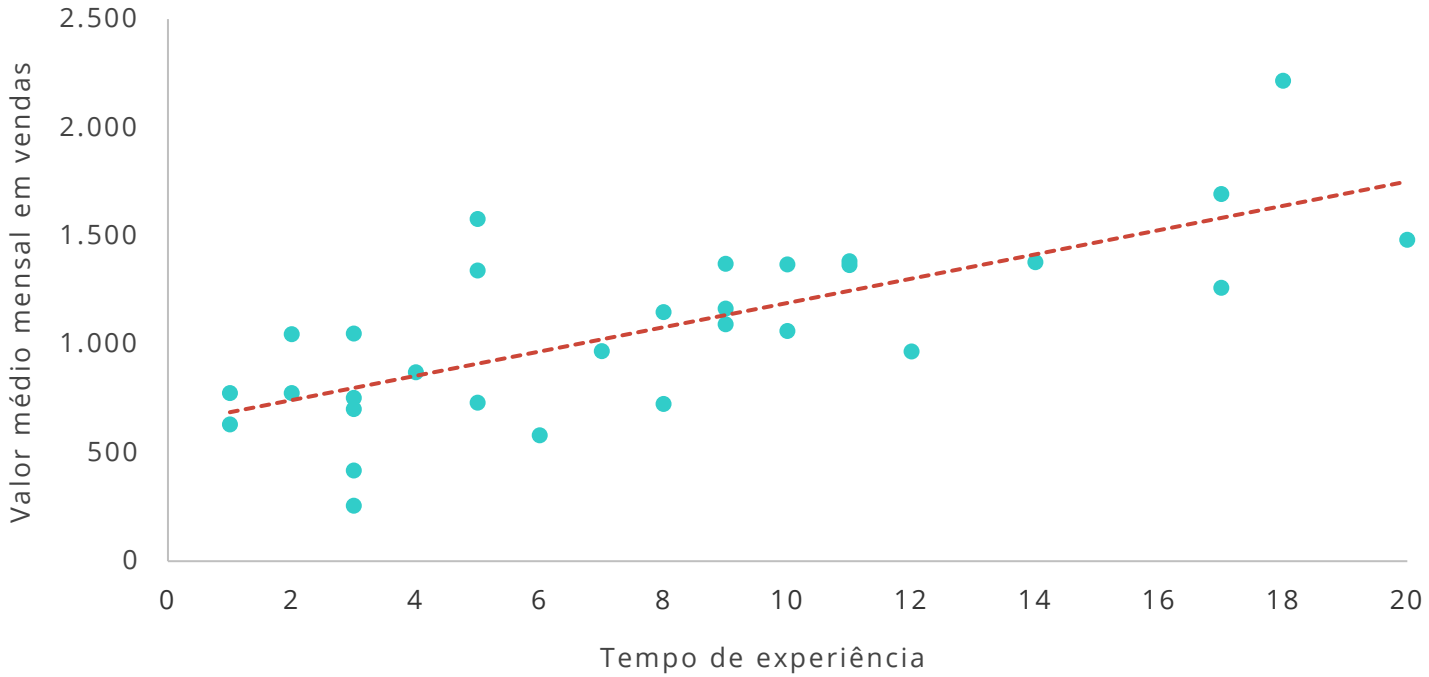
- **Excel**: *CORREL(col1, col2)*
- **R**: *cor(var1, var2)*

Arquivos: Venda_Veiculos (.xlsx e .txt)

4. Equação da Reta

Equação da Reta

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

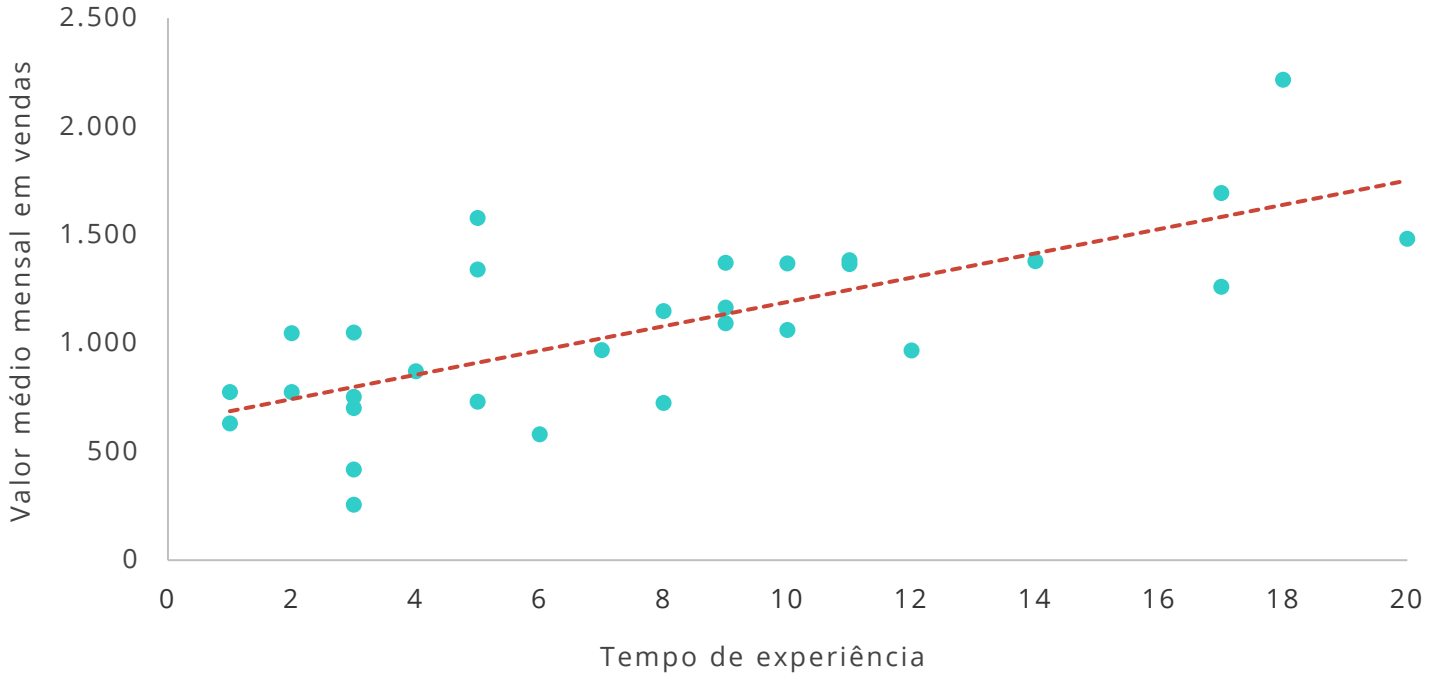


É possível expressar essa relação por meio de alguma regra matemática?

Arquivos: Venda_Veiculos (.xlsx e .txt)

Equação da Reta

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



Equação da reta pontilhada

$$\hat{y} = b_0 + b_1 \cdot x$$

Do ponto de vista matemático, essa equação nada mais é do que uma **fórmula** atendida por todos os pontos da forma $(x, \hat{y})$ que compõem a reta.

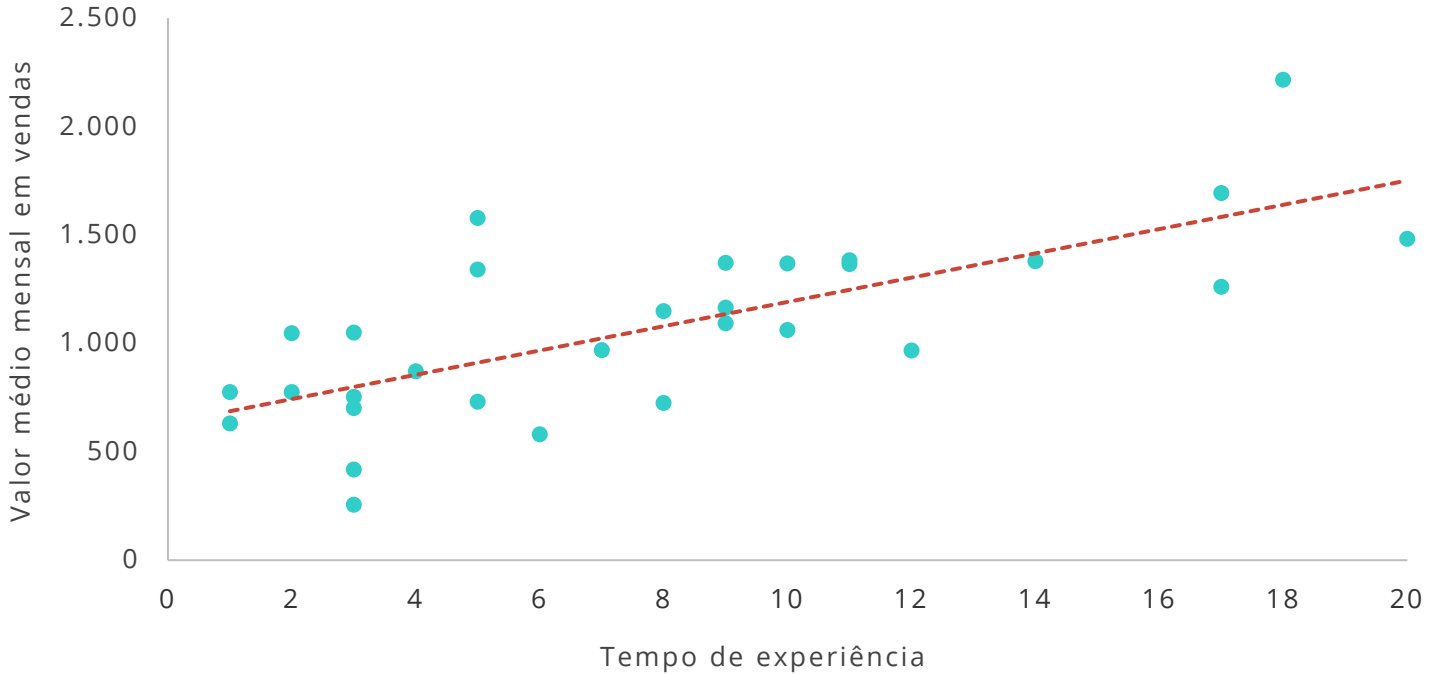
Já do ponto de vista estatístico, essa reta é a que **melhor se ajusta** ao comportamento dos dados.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Equação da Reta

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



Equação da reta pontilhada

$$\hat{y} = b_0 + b_1 \cdot x$$

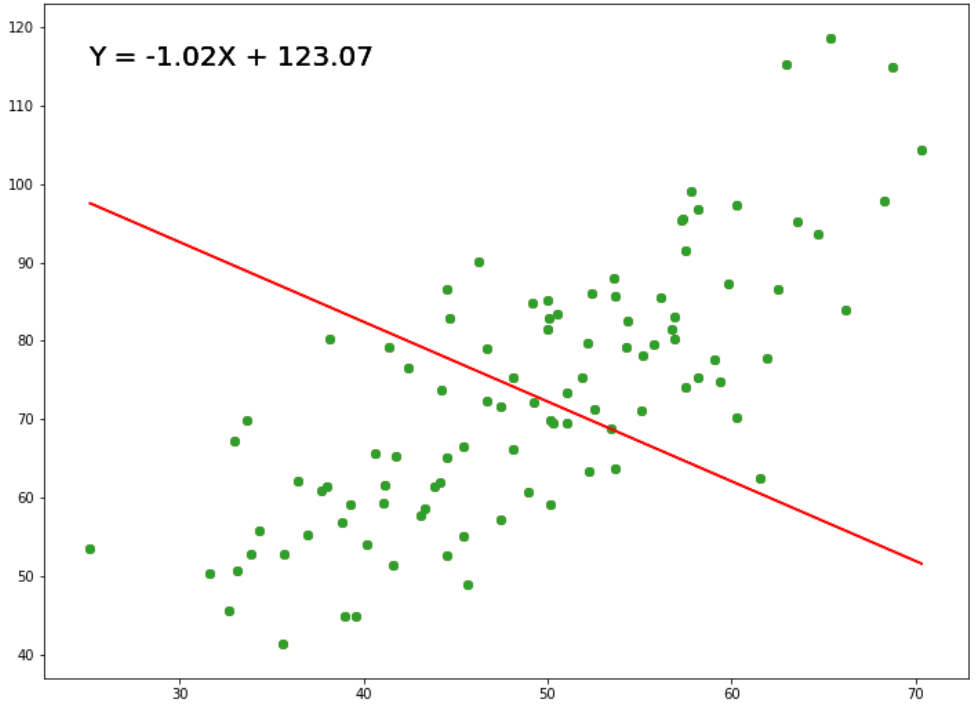
Mas quais são os valores de b_0 e b_1?

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Método de Mínimos Quadrados

A obtenção da reta **ótima**, ou seja, que melhor aproxima a nuvem de pontos, é realizada por meio do método de **mínimos quadrados**. Este método minimiza as **distâncias** entre os pontos e a reta.



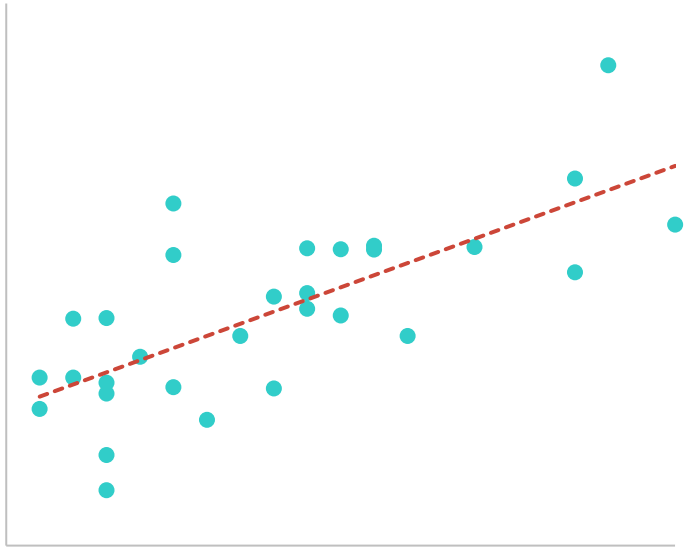
Método de Mínimos Quadrados

Encontrar a reta ótima por meio deste método é equivalente a encontrar os **melhores valores** para b_0 e b_1 na equação.

BUSINESS SCHOOL lab.data

Interpretação da Equação da Reta

Do ponto de vista estatístico, a reta identificada pelo método dos mínimos quadrados é vista como uma simplificação do padrão demonstrado dados observados. É comum chamá-la de **reta ajustada**.

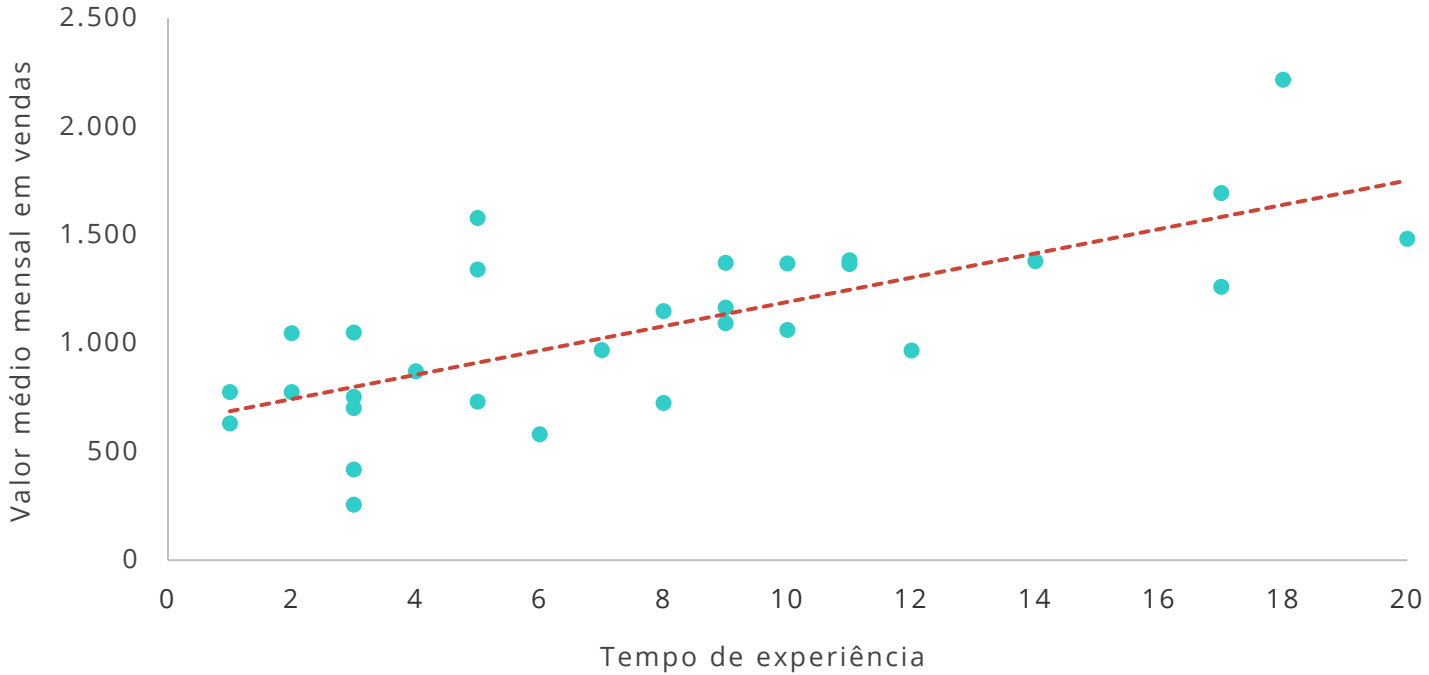


$$\hat{y} = b_0 + b_1 x$$

- ✓ y é o **valor** de uma **variável resposta quantitativa**
- ✓ x é o **valor** de uma **variável explicativa** (por ora, quantitativa)
- ✓ $\hat{y}$ é o **valor ajustado** para a resposta, associado ao valor x
- ✓ b_0 e b_1 são os **coeficientes da reta ajustada**, sendo que:
  - b_0 é o ***intercepto***, que corresponde ao valor ajustado de y quando x = 0.
  - b_1 é o ***ângulo***, que corresponde à variação média no valor ajustado de y quando x aumenta em 1 unidade.

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



$$\hat{y} = 631 + 56 \cdot x$$

Aqui, $b_0 = 631$ e $b_1 = 56$.

O que representa b_0?

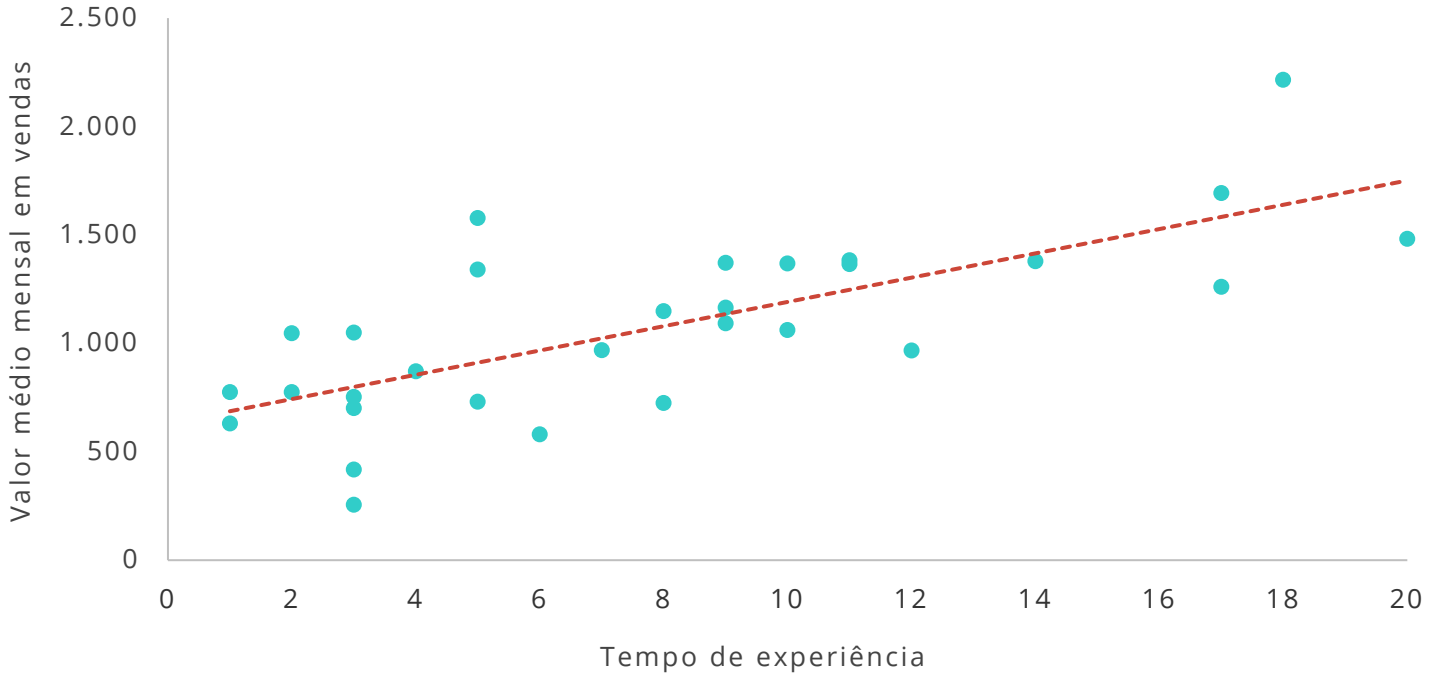
O que representa b_1?

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



$$\hat{y} = 631 + 56 \cdot x$$

Aqui, $b_0 = 631$ e $b_1 = 56$.

O que representa b_0?
O valor médio mensal em vendas para um vendedor que não possua experiência (hipoteticamente) é de **631 mil reais**.

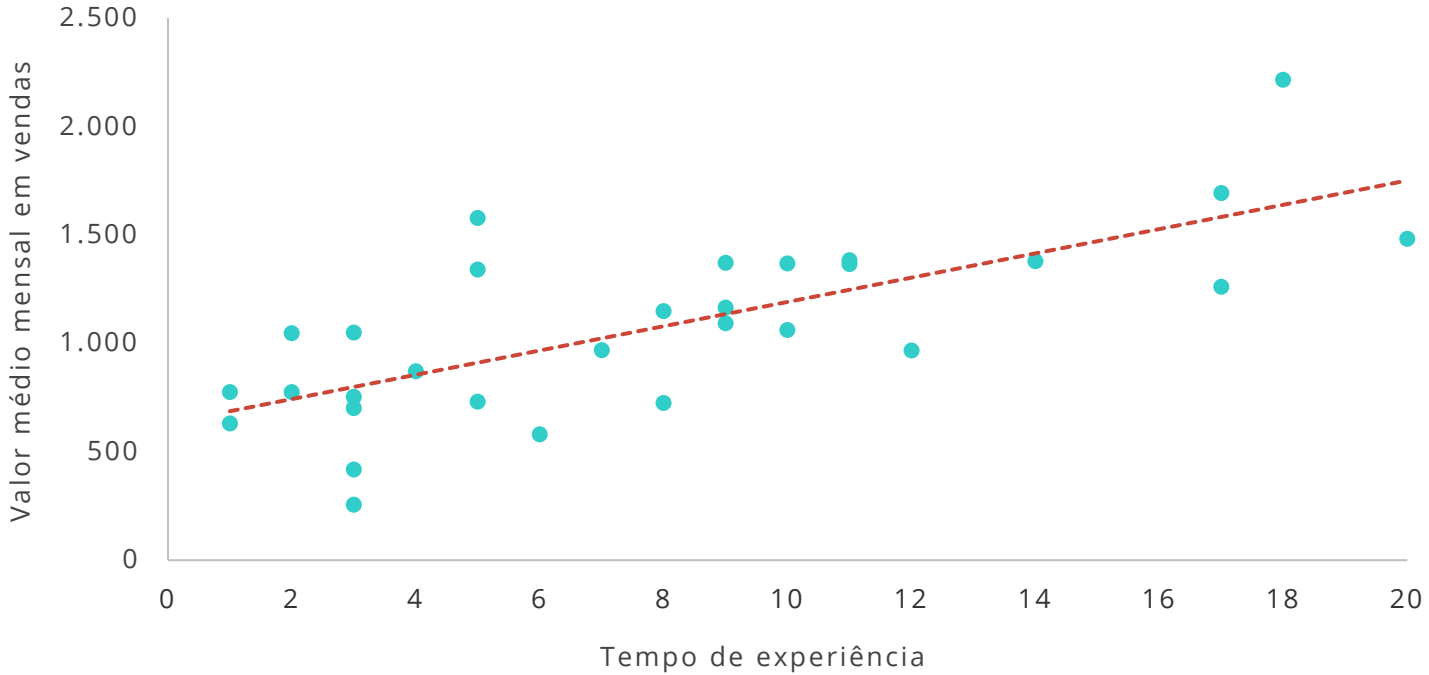
O que representa b_1?
A cada 1 ano a mais de experiência, os vendedores apresentam, em média, **56 mil reais a mais** em vendas por mês.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



$$\hat{y} = 631 + 56 \cdot x$$

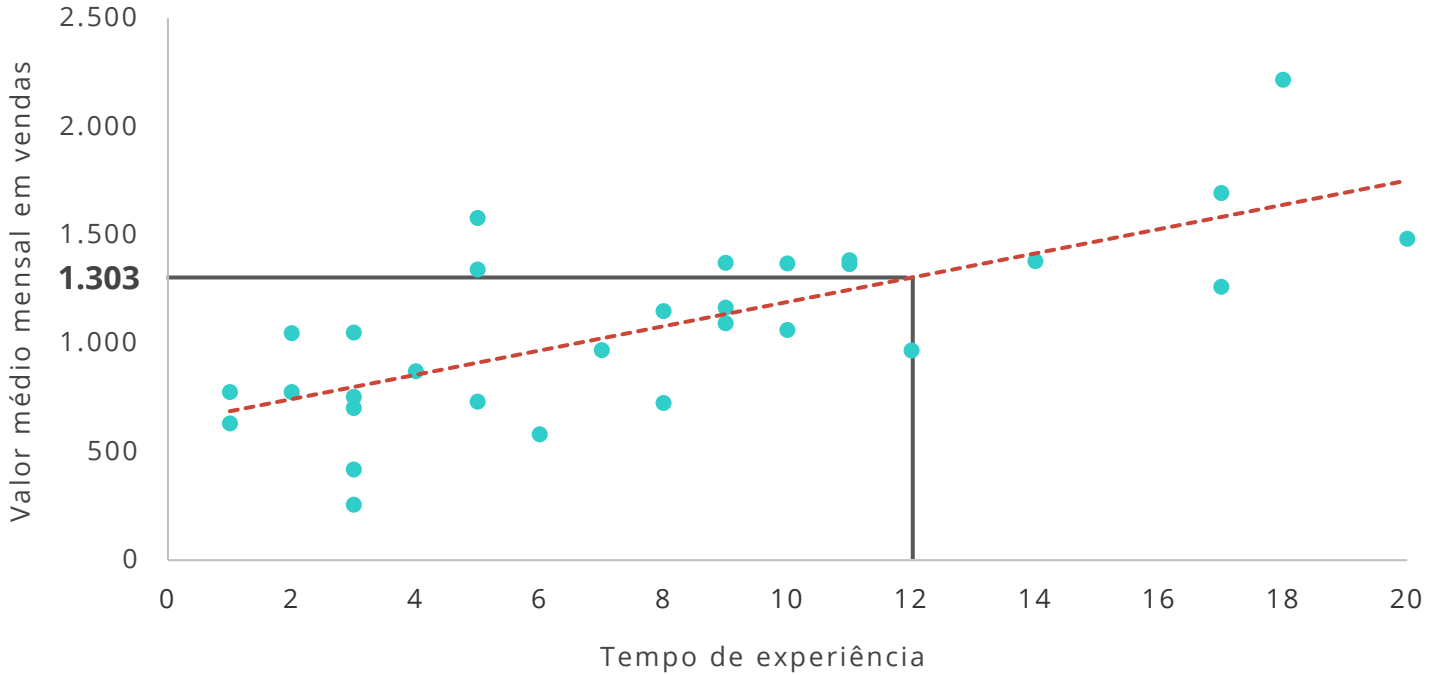
Aqui, $b_0 = 631$ e $b_1 = 56$.

Qual é o valor médio mensal em vendas **ajustado** para um vendedor que possui **12** anos de experiência?

Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



$$\hat{y} = 631 + 56 \cdot x$$

Aqui, $b_0 = 631$ e $b_1 = 56$.

Qual é o valor médio mensal em vendas **ajustado** para um vendedor que possui **12** anos de experiência?

Resp.: 631 + 56 · 12 = **1.303.000** reais

Arquivos: Venda_Veiculos (.xlsx e .txt)

Resíduos

Tendo em vista que a reta é uma aproximação do comportamento observado, os valores $\hat{y}$ obtidos a partir dela tratam-se de **estimativas**, logo, sujeitas a **erros**. Esses erros são denominados **resíduos**, pois tratam-se de informações restantes ("residuais") que não puderam ser explicadas por meio da reta.

Tempo de experiência (em anos)	Valor médio mensal das vendas (em milhares de R$)	**Estimativa** (em milhares de R$)	**Resíduo** (em milhares de R$)
1	775	687	**+88**
1	630	687	**-57**
2	775	743	**+32**
2	1.046	743	**+303**
3	752	799	**-47**
3	255	799	**-544**
3	1.049	799	**+250**
3	701	799	**-98**
3	418	799	**-381**
...	...	...	...

Resíduo =
Resposta real – Resposta estimada

$$e = y - \hat{y}$$

FIA BUSINESS SCHOOL lab.data

Resíduos

Tendo em vista que a reta é uma aproximação do comportamento observado, os valores $\hat{y}$ obtidos a partir dela tratam-se de **estimativas**, logo, sujeitas a **erros**. Esses erros são denominados **resíduos**, pois tratam-se de informações restantes ("residuais") que não puderam ser explicadas por meio da reta.

Tempo de experiência (em anos)	Valor médio mensal das vendas (em milhares de R$)	**Estimativa** (em milhares de R$)	**Resíduo** (em milhares de R$)
1	775	687	**+88**
1	630	687	**-57**
2	775	743	**+32**
2	1.046	743	**+303**
3	752	799	**-47**
3	255	799	**-544**
3	1.049	799	**+250**
3	701	799	**-98**
3	418	799	**-381**
...	...	...	**...**

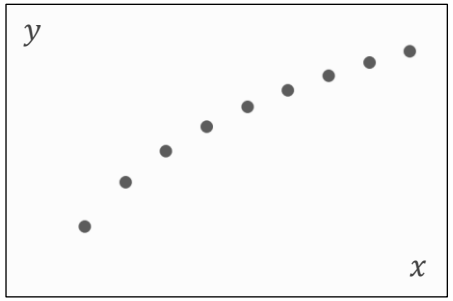
- ✓ A **média dos resíduos** é igual a **zero**, pois o método de mínimos quadrados encontra a reta que passa exatamente pelo região "média" dos pontos, neutralizando desvios positivos e desvios negativos.
- ✓ Dessa forma, deve-se considerar **outras medidas** para julgar a magnitude dos erros (veremos adiante).

Resíduos (tópico extra)

Quando a relação entre as variáveis x e y **não é linear**, isso pode afetar de forma significativa os resíduos. Nesses casos, convém transformar as variáveis de forma a **linearizar a relação**.

Exemplos comuns:

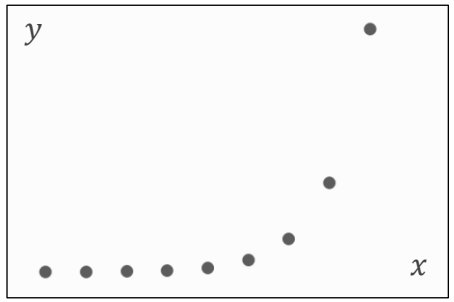
RELAÇÃO LOGARÍTMICA



$$y = a + b \cdot \log(x)$$

Sugestão:
Substituir x por $\exp(x)$, pois $\exp$ é a função inversa do log

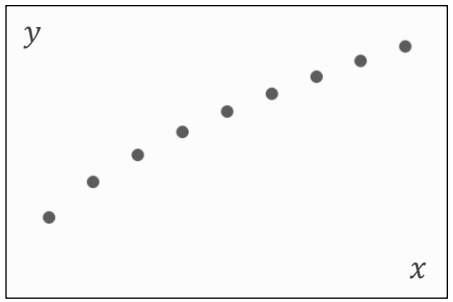
RELAÇÃO EXPONENCIAL



$$y = a + b \cdot \exp(x)$$

Sugestão:
Substituir x por $log(x)$, pois log é a função inversa do exp

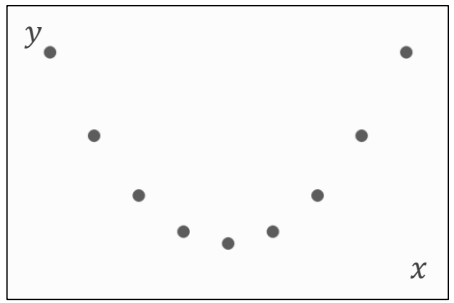
RELAÇÃO RADICIAL



$$y = a + b \cdot \sqrt{x}$$

Sugestão:
Substituir x por x^2, pois o quadrado é a função inversa da raiz

RELAÇÃO QUADRÁTICA



$$y = a \cdot x^2 + b \cdot x + c$$

Não há como linearizar facilmente; requer desmembrar em **dois ajustes de reta**

5. Regressão Linear Simples

Contexto Inferencial

Suponhamos que os dados utilizados para construir a reta ajustada são **amostrais**, e que nosso interesse, agora, está em obter estimativas acerca da variável resposta para os elementos da **população**.

No *case* de venda de veículos, isso poderia se configurar nas seguintes situações:

- Dispomos de dados apenas de uma **amostra aleatória** de vendedores, em vez de todos. Então, teríamos interesse em estimar o valor médio mensal em vendas para vendedores sobre os quais não temos dados.

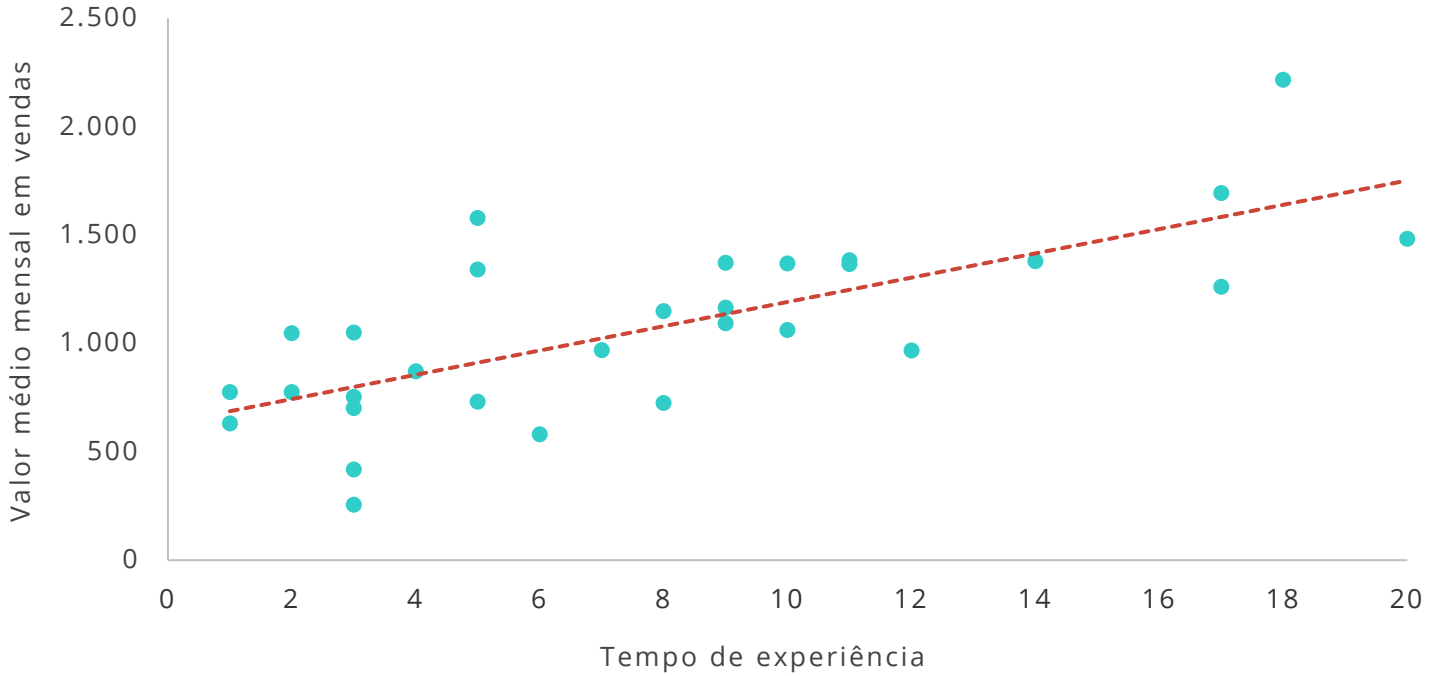
  Essa situação é menos comum num contexto comercial, no qual provavelmente existem dados disponíveis para todos os vendedores, clientes, produtos etc. Porém, é comum em contextos de pesquisas de opinião/satisfação, estudos na área médica e experimentos observacionais em geral.

- Dispomos de dados de **todos** os vendedores atuais, mas queremos estimar o valor médio mensal em vendas para definir metas para **futuros vendedores**, contratados segundo padrões comparáveis aos dos indivíduos que já fazem parte do quadro de funcionários.

  Nesse caso, temos uma amostra temporal, dado que a população completa é composta por elementos intangíveis que serão observados apenas no futuro.

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



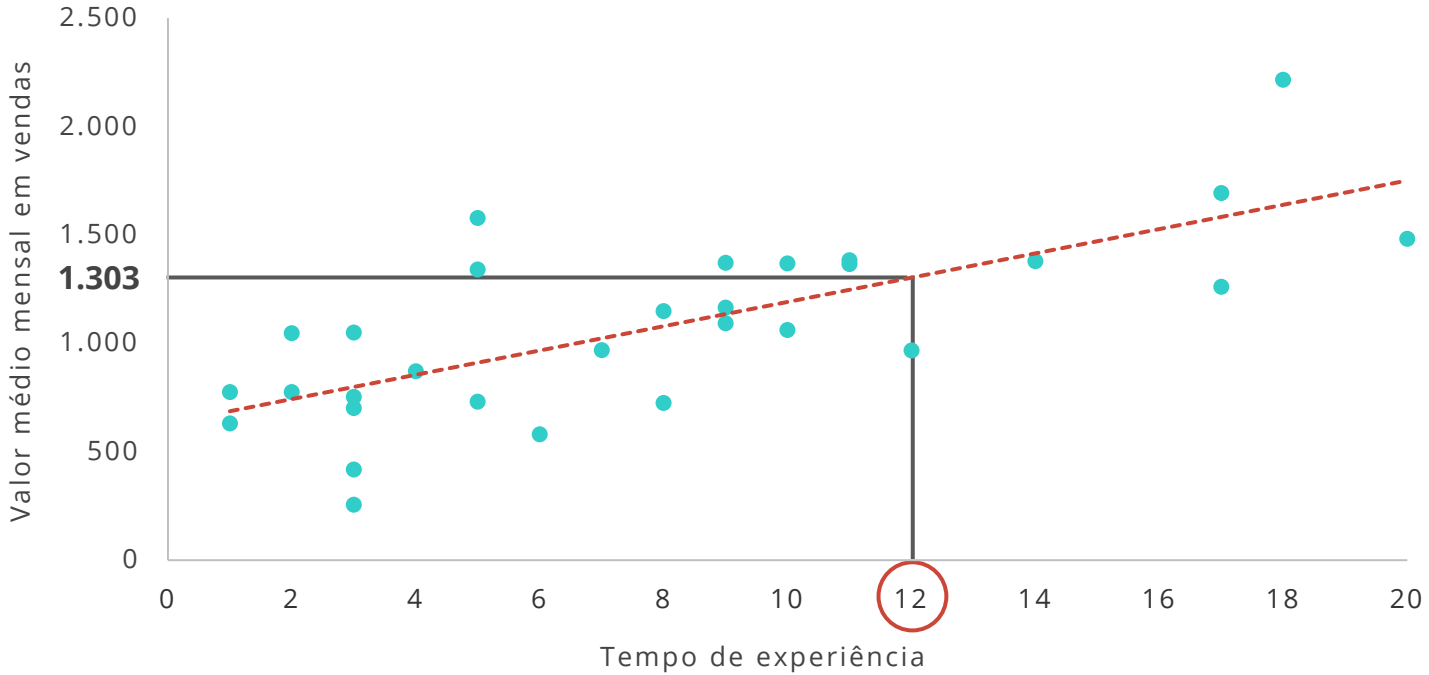
$$\hat{y} = 631 + 56 \cdot x$$

Qual será o valor médio mensal em vendas de um **novo vendedor** que seja contratado com **12** anos de experiência?

Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



$$\hat{y} = 631 + 56 \cdot x$$

Qual será o valor médio mensal em vendas de um **novo vendedor** que seja contratado com **12** anos de experiência?

Não sabemos.

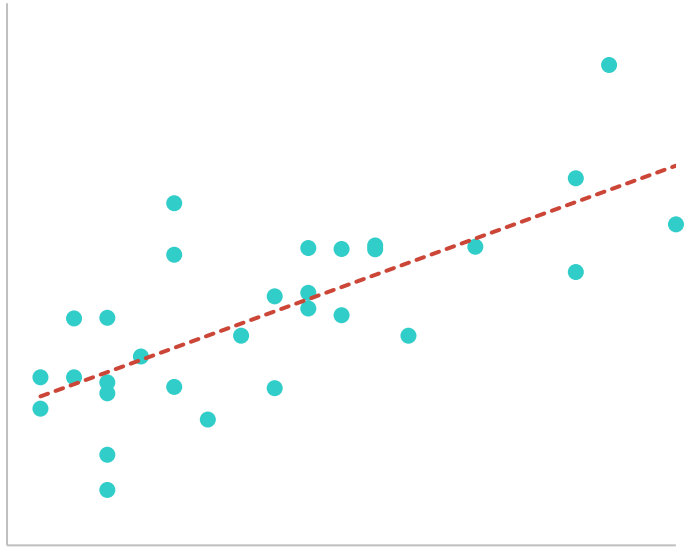
Porém, vamos considerar o valor ajustado de 1.303.000 reais como nossa **estimativa**.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Unificando as Equações

A chave para extrapolação da amostra para a população está em unificar as duas equações que vimos anteriormente, a da **reta ajustada** e a dos **resíduos**, para chegar a uma equação para os **verdadeiros valores** da variável resposta *y*.



Equação da reta ajustada: $\hat{y} = b_0 + b_1 x$

Equação dos resíduos: $e = y - \hat{y}$ ➡ $y = \hat{y} + e$

Equação para y: $y = b_0 + b_1 x + e$

Modelo de Regressão Linear Simples

Podemos, então, estabelecer a equação do **modelo estatístico** de **regressão linear simples**:

Modelo estimado

(a partir de dados da amostra)

$$\hat{y} = b_0 + b_1 x$$

$$y = b_0 + b_1 x + e$$

- ✓ $\hat{y}$ é o **valor estimado da resposta quantitativa**, associado ao valor da variável explicativa x
- ✓ b_0 e b_1 são os **parâmetros estimados**
- ✓ y é o **valor real da resposta**, observado na amostra
- ✓ e é o **erro/resíduo observado**, associado à estimativa

Modelo teórico

(que infere resultados para a população)

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

- ✓ Y é o **valor real da resposta quantitativa**, associado ao valor da variável explicativa X
- ✓ β_0 e β_1 são **parâmetros populacionais desconhecidos**, cujas estimativas correspondem a b_0 e b_1, respectivamente
- ✓ ε é o **erro/resíduo aleatório**, associado à predição de Y a partir de X e dos parâmetros β_0 e β_1

Modelo de Regressão Linear Simples

Podemos, então, estabelecer a equação do **modelo estatístico** de **regressão linear simples**:

Modelo estimado

(a partir de dados da amostra)

$$\hat{y} = b_0 + b_1 x$$

Na estatística inferencial, nosso objetivo é estimar características populacionais, tal como fizemos em aulas anteriores para a média (μ) e a proporção *(p)*.

Agora, nosso objetivo está em estimar os coeficientes populacionais do modelo de regressão linear teórico, β_0 e β_1, e isso será feito a partir das estimativas amostrais b_0 e b_1.

Modelo teórico

(que infere resultados para a população)

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

- ✓ Y é o **valor real da resposta quantitativa**, associado ao valor da variável explicativa X
- ✓ β_0 e β_1 são **parâmetros populacionais desconhecidos**, cujas estimativas correspondem a b_0 e b_1, respectivamente
- ✓ ε é o **erro/resíduo aleatório**, associado à predição de Y a partir de X e dos parâmetros β_0 e β_1

Intervalo de Confiança para β_0 e β_1

Por se tratarem de estimativas amostrais, os valores b_0 e b_1 possuem uma **distribuição** e uma **margem de erro**.

A partir do Teorema do Limite Central, é possível provar que a sua distribuição é aproximadamente **normal**, quando a amostra é grande ($n \geq 30$) e são **não viesados**, ou seja, em média, acertam o verdadeiros valores de β_0 e β_1.

Consequentemente, podemos construir **intervalos de confiança** para β_0 e β_1.

Intervalos de confiança para β_0 e β_1

$$IC(\beta_0; 95\%) = [\, b_0 \pm 1{,}96 \cdot DP(b_0) \,]$$

$$IC(\beta_1; 95\%) = [\, b_1 \pm 1{,}96 \cdot DP(b_1) \,]$$

Ou seja, os intervalos de confiança de β_0 e β_1 estão **centrados nas estimativas amostrais** b_0 e b_1, acrescidas/subtraídas de uma margem de erro.

A magnitude da margem de erro depende do **desvio padrão** (DP) das estimativas, que, por sua vez, é influenciado pelo desvio padrão dos **resíduos**.

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

Apesar de ser possível obter as estimativas b_0 e b_1 a partir do Excel, o **R** fornecerá bem mais insumos para a análise de resultados do modelo de regressão linear simples, usando a função ***lm*** (de *linear model).*

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                 Estimate Std. Error
(Intercept)        630.64      94.96
Tempo_Experiencia   56.01      10.04
```

Algumas medidas resumo dos resíduos

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                 Estimate Std. Error
(Intercept)        630.64      94.96
Tempo_Experiencia   56.01      10.04
```

Estimativas amostrais b_0 e b_1 dos respectivos parâmetros populacionais β_0 e β_1.

Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                 Estimate Std. Error
(Intercept)        630.64      94.96
Tempo_Experiencia   56.01      10.04
```

Desvios padrão amostrais de b_0 e b_1

Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                  Estimate Std. Error
(Intercept)         630.64      94.96
Tempo_Experiencia    56.01      10.04
```

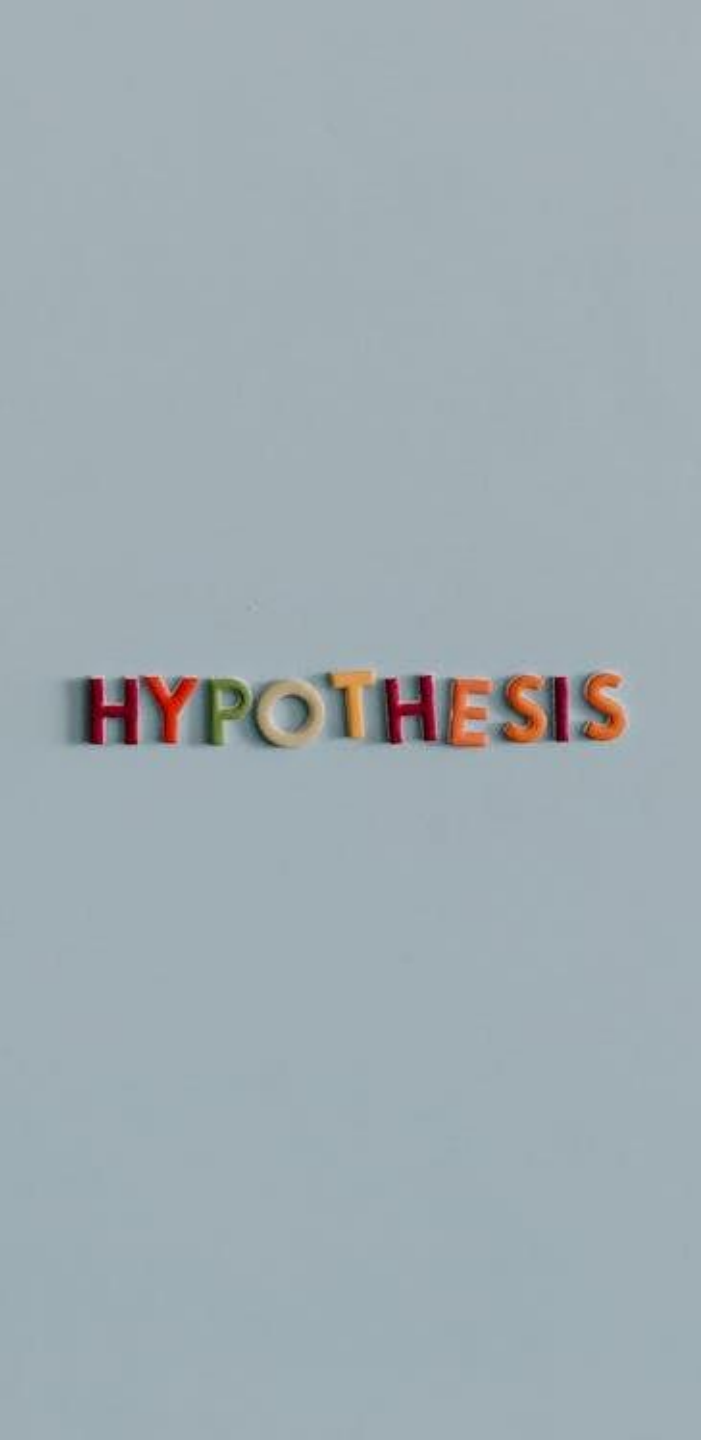
$$IC(\beta_0; 95\%) = [\, b_0 \pm 1{,}96 \cdot DP(b_0) \,]$$
$$= [\, 630{,}64 \pm 1{,}96 \cdot 94{,}96 \,]$$
$$= [\, 444{,}52;\ 816{,}76]$$

$$IC(\beta_1; 95\%) = [\, b_1 \pm 1{,}96 \cdot DP(b_1) \,]$$
$$= [\, 56{,}01 \pm 1{,}96 \cdot 10{,}04 \,]$$
$$= [\, 36{,}33;\ 75{,}69]$$

O que os intervalos de confiança nos dizem de importante?

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

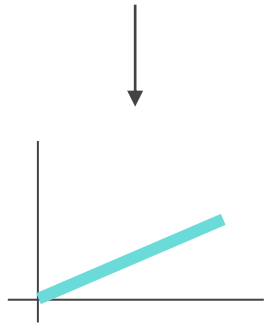


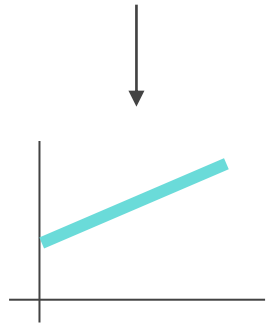
Hipóteses de Interesse

Um particular conjunto de **hipóteses de interesse** a respeito do intercepto populacional β_0 são:

$$\beta_0 = 0 \quad \text{ou} \quad \beta_0 \neq 0$$

Reta cruza o eixo y na origem do plano cartesiano, ou seja, no ponto (0,0)





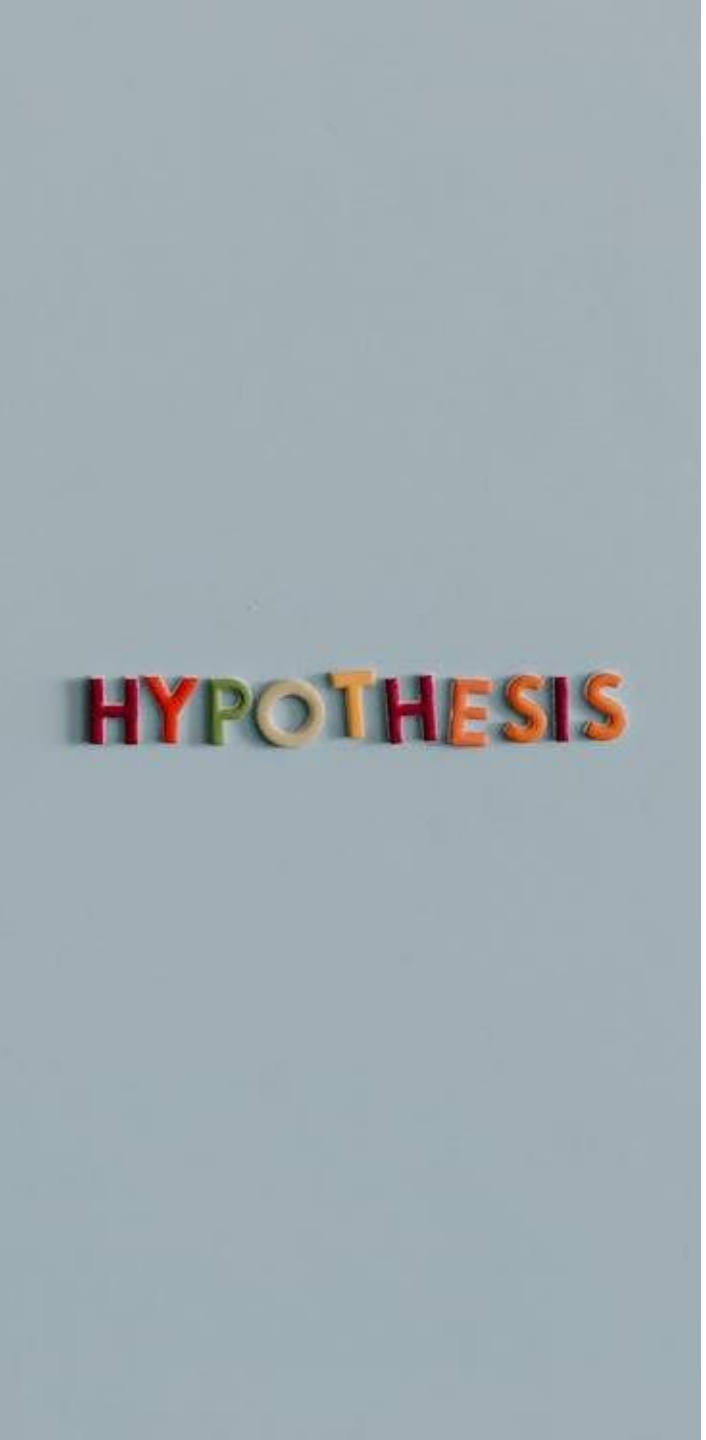
Reta cruza o eixo y em outro ponto, que não a origem (0,0)

Caso tenhamos alta confiança para afirmar que $\boldsymbol{\beta_0 = 0}$, podemos simplificar a equação do modelo para:

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

Este é o **princípio da parcimônia**: explicar um fenômeno a partir da menor quantidade possível de componentes/fatores.

FIA BUSINESS SCHOOL lab.data



Hipóteses de Interesse

Um particular conjunto de **hipóteses de interesse** a respeito do intercepto populacional β_1 são:

$$\beta_1 = 0 \quad \text{ou} \quad \beta_1 \neq 0$$

Reta tem ângulo de inclinação nulo, ou seja, é uma reta horizontal

Reta tem ângulo de inclinação não nulo, ou seja, possui uma tendência

Caso tenhamos alta confiança para afirmar que $\boldsymbol{\beta_1 = 0}$, a variável $\boldsymbol{X}$ perde a relevância e o modelo se reduz a:

$$Y = \beta_0 + \varepsilon$$

Esta é a principal **questão de negócio** que queremos responder: existe, de fato, relação linear entre as variáveis $\boldsymbol{X}$ e $\boldsymbol{Y}$?

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

$$IC(\beta_0; 95\%) = [\, b_0 \pm 1{,}96 \cdot DP(b_0) \,]$$
$$= [\, 630{,}64 \pm 1{,}96 \cdot 94{,}96 \,]$$
$$= [\, \mathbf{444{,}52;\ 816{,}76} \,]$$

$$IC(\beta_1; 95\%) = [\, b_1 \pm 1{,}96 \cdot DP(b_1) \,]$$
$$= [\, 56{,}01 \pm 1{,}96 \cdot 10{,}04 \,]$$
$$= [\, \mathbf{36{,}33;\ 75{,}69} \,]$$

Como os intervalos de confiança de β_0 e β_1 **não abrangem o valor zero**, podemos concluir, com **95% de confiança**, que:

- Não é possível simplificar o modelo retirando o intercepto.
- O tempo de experiência dos vendedores possui relação linear **estatisticamente significativa** com o valor médio mensal em vendas.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Abordagem por Teste de Hipóteses

Uma alternativa à construção do intervalo de confiança é a realização de um **teste de hipóteses**.

As duas abordagens são **equivalentes**, mas os testes de hipóteses serão úteis para nós daqui em diante.

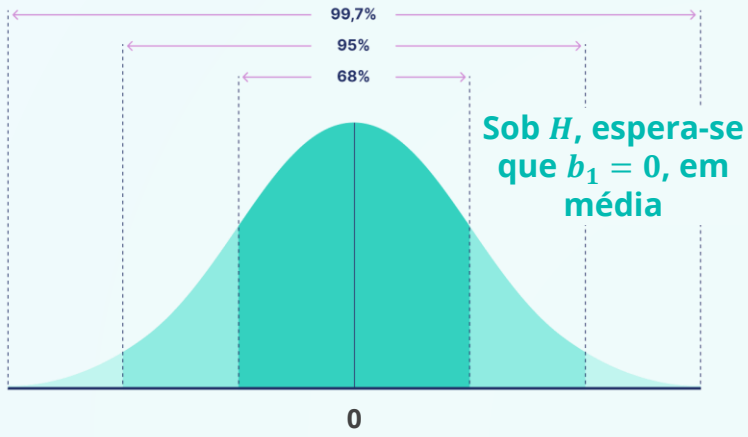
Racional do teste de hipóteses para $\boldsymbol{\beta_1}$:

1. Estabelecemos duas hipóteses:

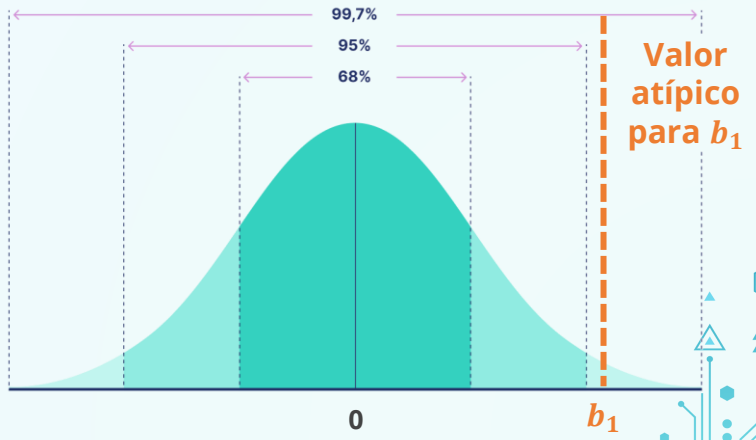
$$H\colon\ \beta_1 = 0$$
$$A\colon\ \beta_1 \neq 0$$

H é chamada ***hipótese nula*** e A é chamada ***hipótese alternativa***.

2. Sabemos que a distribuição teórica de b_1 é aproximadamente normal, com média β_1. **Supondo** que H seja verdadeira, tal distribuição seria:



3. Ao coletar uma amostra, avaliamos o valor obtido de b_1. Se estiver muito distante de zero, isso **contradiz** a afirmação de que H é verdadeira:



FIA BUSINESS SCHOOL lab.data

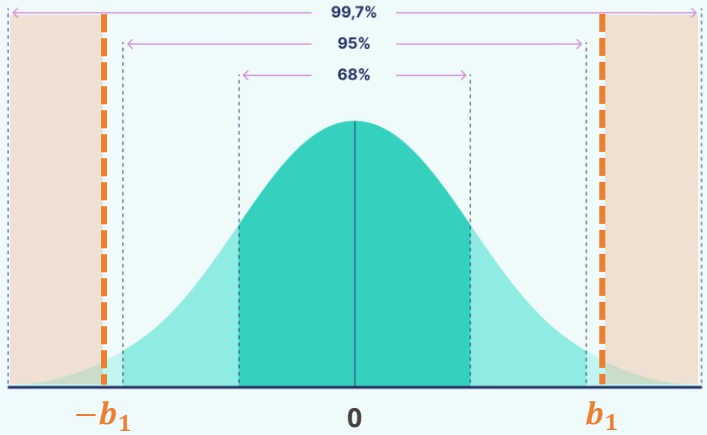
Abordagem por Teste de Hipóteses

Uma alternativa à construção do intervalo de confiança é a realização de um **teste de hipóteses**.

As duas abordagens são **equivalentes**, mas os testes de hipóteses serão úteis para nós daqui em diante.

Racional do teste de hipóteses para $\boldsymbol{\beta_1}$:

4. Calcula-se a **probabilidade** de se observar um valor ainda mais extremo do que o observado para b_1, nas duas caudas da distribuição normal:



5. Essa probabilidade é denominada ***p-valor***, ou ***valor-p***.
A decisão final do teste é:

- Se **p-valor** $\leq$ **5%**, rejeitamos a hipótese H e aceitamos a hipótese A, com confiança de 95%.
- Se **p-valor** $>$ **5%**, não rejeitamos a hipótese H, com confiança de 95%.

O p-valor é útil pois, ao contrário do intervalo de confiança, ele resume em um único valor o **grau de evidência**, ou **"força"**, que se tem para rejeitar ou não a hipótese H.

O processo é **idêntico** para testar a hipótese de nulidade para $\boldsymbol{\beta_0}$.

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)        630.64      94.96   6.641 0.000000332 ***
Tempo_Experiencia   56.01      10.04   5.581 0.000005683 ***
```

p-valores dos testes de nulidade de $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$, em escala de 0 (0%) a 1 (100%).

Arquivos: Venda_Veiculos (.xlsx e .txt)

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                  Estimate Std. Error t value   Pr(>|t|)
(Intercept)         630.64      94.96   6.641 0.000000332 ***
Tempo_Experiencia    56.01      10.04   5.581 0.000005683 ***
```

Como ambos os *p*-valores são pequenos (inferiores a 0,05 ou 5%), então, com **95% de confiança**:

- Rejeitamos a hipótese H: $\beta_0 = 0$, o que indica que não é possível simplificar o modelo retirando o intercepto.
- Rejeitamos a hipótese H: $\beta_1 = 0$, o que indica que o tempo de experiência dos vendedores possui relação linear **estatisticamente significativa** com o valor médio mensal em vendas.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Diagnóstico do Modelo

A fase final de construção de um modelo de regressão linear é a de diagnóstico, na qual avaliamos a **qualidade do ajuste** e algumas propriedades dos **resíduos**.

Qualidade de ajuste

A variável explicativa traz bastante informação sobre a variável resposta?

- Coeficiente de determinação *(R^2)*
- Erro absoluto médio *(MAE)*
- Erro absoluto médio percentual *(MAPE)*

Propriedades dos resíduos

Os erros/resíduos se comportam de forma apropriada?

- Normalidade
- Homocedasticidade (variância constante *versus* X)
- Independência

Qualidade de Ajuste

Coeficiente de determinação (R^2)

Corresponde ao percentual de **comportamento explicado** da variável resposta por meio da variável explicativa. Pode variar entre 0% e 100%.

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Note que:

- Numa situação extrema em que os valores x não explicam **nada** sobre y, todos os valores preditos $\hat{y}$ seriam iguais à média geral, ou seja, $\hat{y} = \bar{y}$. Então, o numerador torna-se igual a zero, e portanto, $R^2 = 0 = 0\%$.
- No extremo oposto em que os valores x explicam y **perfeitamente**, todos os valores preditos $\hat{y}$ seriam iguais ao valor real, ou seja, $\hat{y} = y$. Então, o numerador torna-se igual ao denominador, e portanto, $R^2 = 1 = 100\%$.

Qualidade de Ajuste

Coeficiente de determinação (R^2)

Na regressão linear simples, o valor de R^2 coincide com o **quadrado** do **coeficiente de correlação linear** *(r)*. Ou seja, $R^2 = r^2$.

Sugestão de interpretação

Valor	Explicabilidade
$R^2 \geq 0{,}80$	Muito boa
$0{,}60 \leq R^2 < 0{,}80$	Boa
$0{,}40 \leq R^2 < 0{,}60$	Moderada
$0{,}20 \leq R^2 < 0{,}40$	Baixa
$R^2 < 0{,}20$	Muito baixa

Tal como a correlação linear, a interpretação da força de explicabilidade por meio do R^2 é **subjetiva** e pode variar a depender do contexto.

Qualidade de Ajuste

Erro absoluto médio *(MAE)* e erro absoluto médio percentual *(MAPE)*

Correspondem à **média absoluta dos resíduos**, em sua própria escala ou em percentual, sendo este relativo ao patamar da variável resposta y.

$$MAE = \frac{\sum_i |e_i|}{n}$$

$$MAPE = \frac{\sum_i |e_i / y_i|}{n}$$

É importante avaliar os resíduos de forma absoluta, pois a média dos valores originais é sempre **zero**, em decorrência do método dos mínimos quadrados.

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

```
Call:
lm(formula = Valor_Medio_Mensal_Vendas ~ Tempo_Experiencia, data = dados_vendas)

Residuals:
    Min      1Q  Median      3Q     Max
-543.66 -167.94  -10.21  131.77  667.32

Coefficients:
                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)        630.64      94.96   6.641 0.000000332 ***
Tempo_Experiencia   56.01      10.04   5.581 0.000005683 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 289 on 28 degrees of freedom
Multiple R-squared:  0.5266,    Adjusted R-squared:  0.5097
```

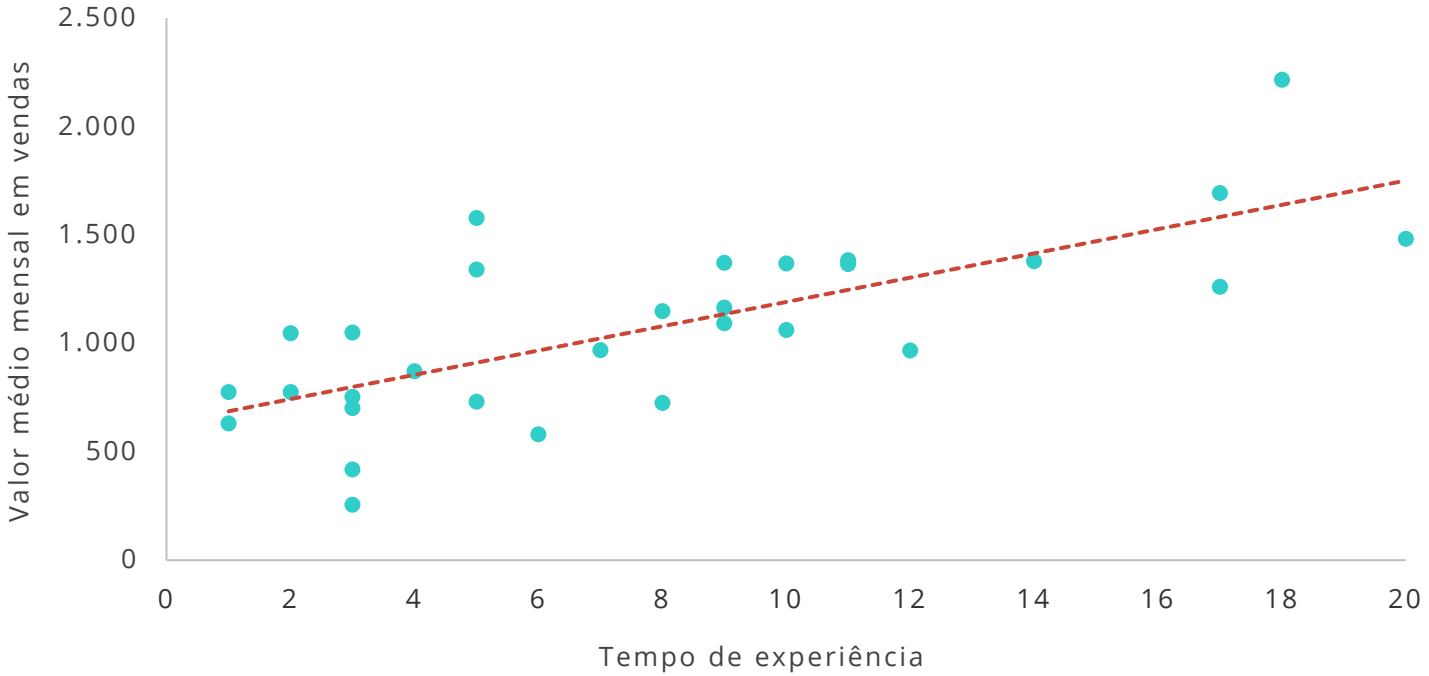
Coeficiente de determinação (R^2) do modelo

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?



R^2 = 0,5266

Cerca de 53% do comportamento do valor médio mensal em vendas de um vendedor pode ser explicado a partir do seu tempo de experiência.

***MAE* = 216**
***MAPE* = 27%**

Em média, o modelo erra o valor médio mensal de vendas em 216 mil reais, para mais ou para menos. Isso corresponde a uma média de 27% de erro em relação ao valor almejado.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

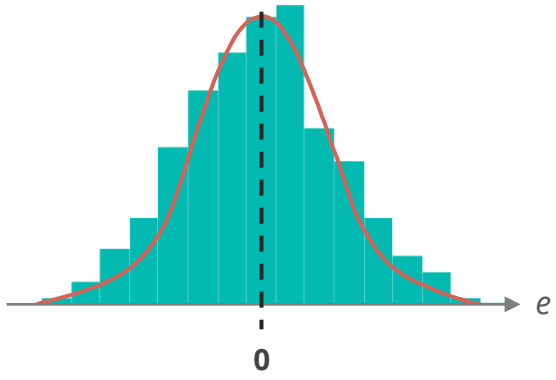
Propriedades dos Resíduos

Normalidade

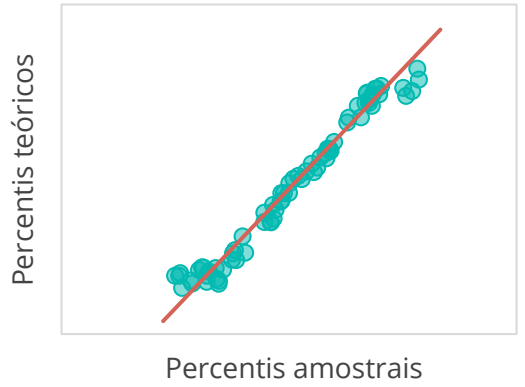
Os resíduos devem seguir uma **distribuição normal**, ou seja, com alta concentração em torno da sua média (zero) e decaimento simétrico para ambos os lados.

Isso pode ser avaliado por meio de gráficos como o **histograma** e o ***Q-Q plot***.

Histograma dos resíduos



Q-Q plot dos resíduos



O ***Q-Q plot*** é um gráfico de resíduos que relaciona os percentis dos resíduos obtidos no ajuste do modelo (na horizontal) *versus* os percentis teóricos esperados sob uma distribuição normal (na vertical).

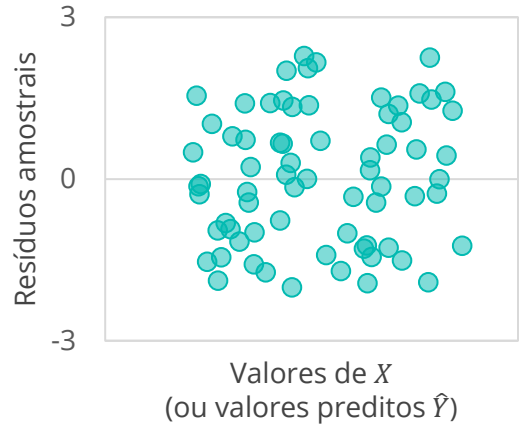
Propriedades dos Resíduos

Homocedasticidade

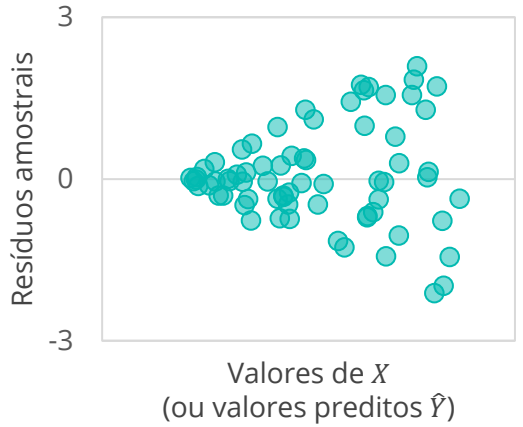
Os resíduos devem ter **variabilidade constante** para qualquer valor de X. Ou seja, a magnitude de erro do modelo não pode depender do valor da variável explicativa.

Isso pode ser avaliado por meio de um **gráfico de dispersão** entre os resíduos do modelo e os valores da variável X. Espera-se não encontrar nenhum padrão de comportamento específico, e sim aleatoriedade.

Exemplo de resíduos com variabilidade **constante**



Exemplo de resíduos com variabilidade **crescente**



FIA BUSINESS SCHOOL lab.data

Propriedades dos Resíduos

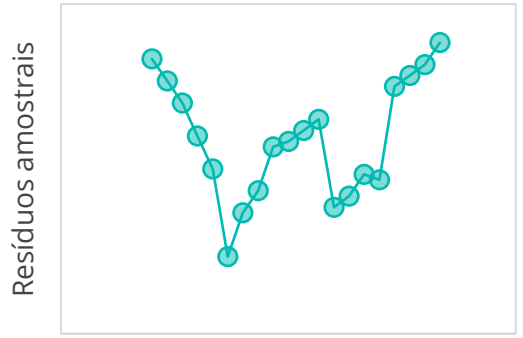
Independência

Os resíduos devem ser **independentes** entre si. Isso é equivalente a afirmar que a resposta manifestada por uma observação (e consequentemente, o erro) não interfere na resposta manifestada por outra observação.

Não há uma forma simples de verificar a independência, pois este é um conhecimento prévio que devemos ter acerca da natureza da nossa amostra.

O exemplo mais preocupante de **erros dependentes** surge quando as observações representam instantes de tempo (série histórica).

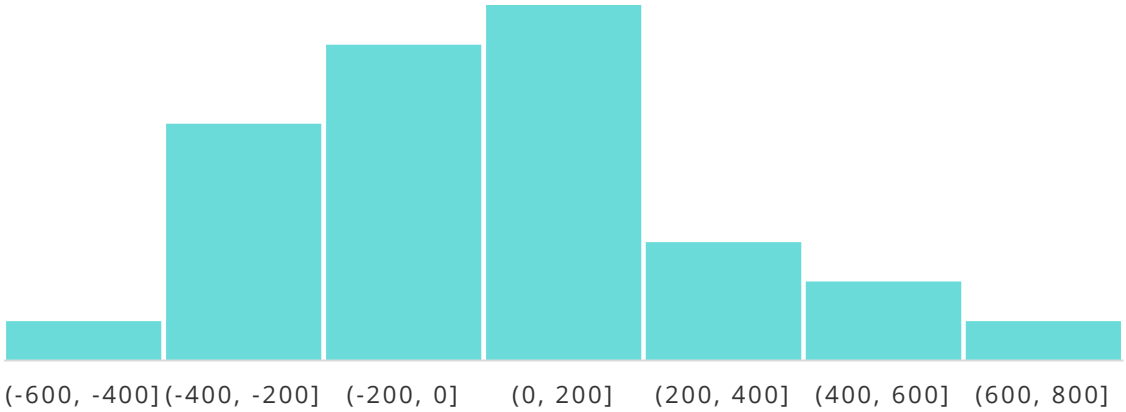
Exemplo de resíduos para **dados de série histórica**



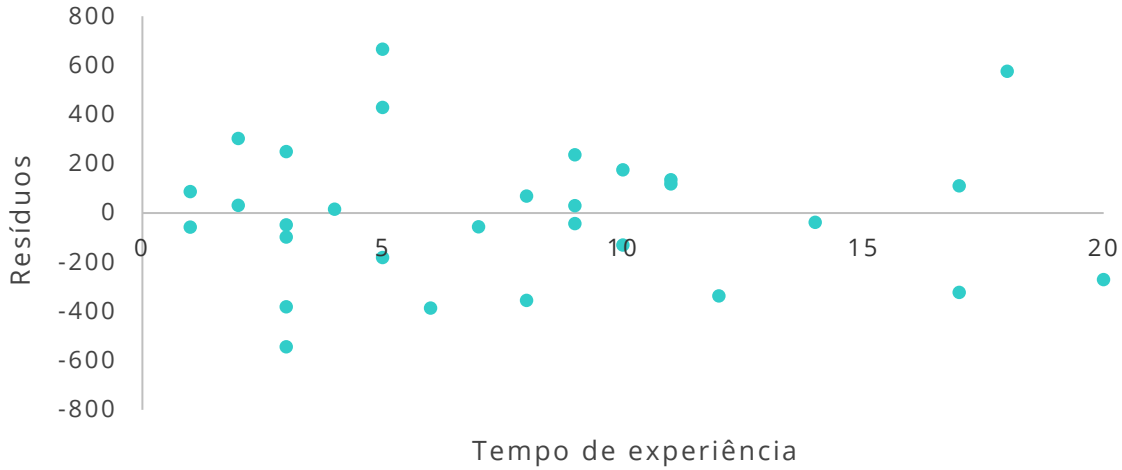
Case: Venda de Veículos

Uma concessionária de veículos deseja estimar o valor médio mensal (R$) em vendas de veículos, em função do tempo de experiência dos vendedores. Existe relação linear entre esses dois aspectos?

Resíduos



Resíduos *versus* tempo de experiência



Os resíduos aparentam ter uma distribuição levemente assimétrica à direita, em vez de uma distribuição normal; ou seja, os valores reais (y) tendem a ser **inferiores** aos valores preditos ($\hat{y}$).

Apesar disso, parece haver homocedasticidade: mesmo padrão de variabilidade de erros de acordo com o tempo de experiência.

Arquivos: Venda_Veiculos (.xlsx e .txt)

FIA BUSINESS SCHOOL lab.data

Case: Epidemiologia

Uma agência de saúde tem o objetivo de entender a relação entre pluviometria (nível de chuva) em um mês de verão e o registro de casos de dengue no mês seguinte. Para isso, foram coletados dados pluviométricos de 45 cidades no último mês de janeiro, bem como a quantidade de casos de dengue reportados no mês de fevereiro.

ID_CIDADE	VOLUME_CHUVA_JAN_MM	QTDE_CASOS_DENGUE_FEV
CID_01	212	12779
CID_02	241	12717
CID_03	253	14997
CID_04	262	14022
CID_05	311	15383
...	...	...

Arquivo: Epidemiologia (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Epidemiologia

Uma agência de saúde tem o objetivo de entender a relação entre pluviometria (nível de chuva) em um mês de verão e o registro de casos de dengue no mês seguinte. Para isso, foram coletados dados pluviométricos de 45 cidades no último mês de janeiro, bem como a quantidade de casos de dengue reportados no mês de fevereiro.

a) Faça uma breve análise exploratória da base de dados.
b) De forma gráfica, parece existir relação linear entre o nível de chuva no mês e a quantidade de casos de dengue no mês seguinte? Calcule e interprete o coeficiente de correlação linear entre essas duas variáveis.
c) Construa um modelo de regressão linear simples. Interprete as estimativas dos parâmetros, os intervalos de 95% de confiança e os *p*-valores. Podemos dizer que existe associação linear estatisticamente significativa entre o nível de chuva no mês e a quantidade de casos de dengue no mês seguinte, com 95% de confiança?
d) Escreva a equação estimada do modelo final.
e) Interprete o valor do coeficiente de determinação (R^2). Como você avalia a qualidade do modelo?
f) Analise graficamente os resíduos do modelo. Eles seguem um comportamento razoável?
g) Refaça o gráfico do item (b), acrescentando a reta ótima estimada.
h) Estime a quantidade de casos de dengue em fevereiro para uma cidade que teve 280mm de chuva em janeiro.

Arquivo: Epidemiologia (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Faturamento em *E-commerce*

Uma empresa que comercializa produtos eletrônicos por meio de *e-commerce* deseja compreender se o faturamento bruto (R$) obtido no último mês está associado ao investimento realizado (R$) em anúncios em mídias digitais para o respectivo produto, no mesmo período.

COD_PRODUTO	INVESTIMENTO	FATURAMENTO
C_0001	20.500	526.400
C_0002	17.500	312.200
C_0003	18.500	429.200
C_0004	14.500	470.900
C_0005	12.000	407.100
C_0006	18.800	339.800
C_0007	18.700	335.900
C_0008	15.000	361.300
C_0009	16.700	341.700
C_0010	23.500	484.300
...	...	...

Arquivo: Faturamento (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Faturamento em *E-commerce*

Uma empresa que comercializa produtos eletrônicos por meio de *e-commerce* deseja compreender se o faturamento bruto (R$) obtido no último mês está associado ao investimento realizado (R$) em anúncios em mídias digitais para o respectivo produto, no mesmo período.

a) Faça uma breve análise exploratória da base de dados.
b) De forma gráfica, parece existir relação linear entre o investimento em mídias digitais e o faturamento bruto por produto? Calcule e interprete o coeficiente de correlação linear entre essas duas variáveis.
c) Construa um modelo de regressão linear simples. Interprete as estimativas dos parâmetros, os intervalos de 95% de confiança e os *p*-valores. Podemos dizer que existe associação linear estatisticamente significativa entre o investimento em mídias digitais e o faturamento bruto por produto, com 95% de confiança?
d) Escreva a equação estimada do modelo final.
e) Interprete o valor do coeficiente de determinação (R^2). Como você avalia a qualidade do modelo?
f) Analise graficamente os resíduos do modelo. Eles seguem um comportamento razoável?
g) Refaça o gráfico do item (b), acrescentando a reta ótima estimada.
h) Estime o faturamento bruto para um produto que tenha 18.000 reais de investimento em mídias digitais.

Arquivo: Faturamento (.txt)

FIA BUSINESS SCHOOL lab.data

6. Regressão Linear Múltipla

Modelo de Regressão Linear Múltipla

Podemos estender naturalmente o modelo de regressão linear para o caso de **múltiplas variáveis explicativas**.

Modelo estimado

(a partir de dados da amostra)

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + e$$

- ✓ $\hat{y}$ é o **valor estimado da resposta quantitativa**, associado aos valores das variáveis explicativas $x_1, \ldots, x_k$
- ✓ b_0, b_1, ... , b_k são os **parâmetros estimados**
- ✓ y é o **valor real da resposta**, observado na amostra
- ✓ e é o **resíduo/erro observado**, associado à estimativa

Modelo teórico

(que infere resultados para a população)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

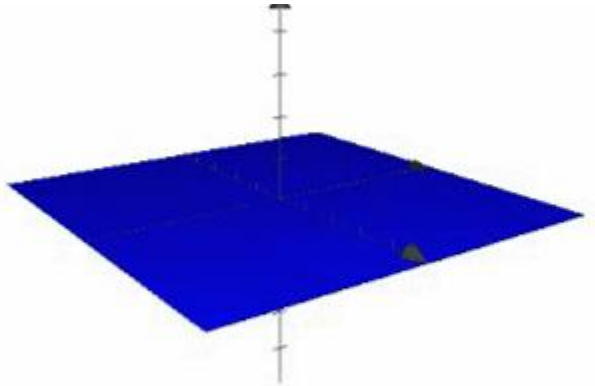
- ✓ Y é o **valor real da resposta quantitativa**, associado aos valores das variáveis explicativas $X_1, \ldots, X_k$
- ✓ β_0, β_1 ... β_k são **parâmetros populacionais desconhecidos**, cujas estimativas correspondem a b_0, b_1, ... , b_k, respectivamente
- ✓ ε é o **resíduo/erro aleatório**, associado à predição de Y a partir de $X_1, \ldots, X_k$ e dos parâmetros β_0, β_1 ... β_k

FIA BUSINESS SCHOOL lab.data

Visualização no Caso Tridimensional

Exemplo: modelo ajustado **tridimensional** (Y, X_1 e X_2)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$



Créditos da imagem: https://commons.wikimedia.org/wiki/File:2d_multiple_linear_regression.gif

Case: Limite de Cartão de Crédito

Uma instituição financeira tem o objetivo de estabelecer uma regra automatizada para definição do valor de **limite de cartão de crédito ideal** para cada novo cliente. Para isso, pretende se basear nos padrões históricos de limites definidos de forma personalizada pelos gerentes, que levavam em conta uma série de fatores em potencial.

Variável	Descrição
ID_CLIENTE	Código identificador do cliente
SCORE_CREDITO	*Score* de crédito do cliente no mercado
VALOR_TOTAL_ATRASOS_12M	Valor total (R$) que o cliente atrasou em pagamentos de cartão de crédito, empréstimos etc., nos últimos 12 meses
QTD_CONSULTAS_CREDITO_12M	Quantidade de consultas de crédito ao nome do cliente, realizadas por outras instituições, nos últimos 12 meses
RENDIMENTO_MEDIO_12M	Rendimento médio mensal do cliente, nos últimos 12 meses
TEMPO_TRABALHO	Tempo de trabalho do cliente, como empregado formal (CLT) ou autônomo
IDADE	Idade do cliente, em anos
QTD_DEPENDENTES	Quantidade de dependentes do cliente
LIMITE_INICIAL_CARTAO	Limite inicial de cartão de crédito definido para o cliente (R$)

Arquivo: Limite_Cartao (.txt)

Case: Limite de Cartão de Crédito

Uma instituição financeira tem o objetivo de estabelecer uma regra automatizada para definição do valor de **limite de cartão de crédito ideal** para cada novo cliente. Para isso, pretende se basear nos padrões históricos de limites definidos de forma personalizada pelos gerentes, que levavam em conta uma série de fatores em potencial.

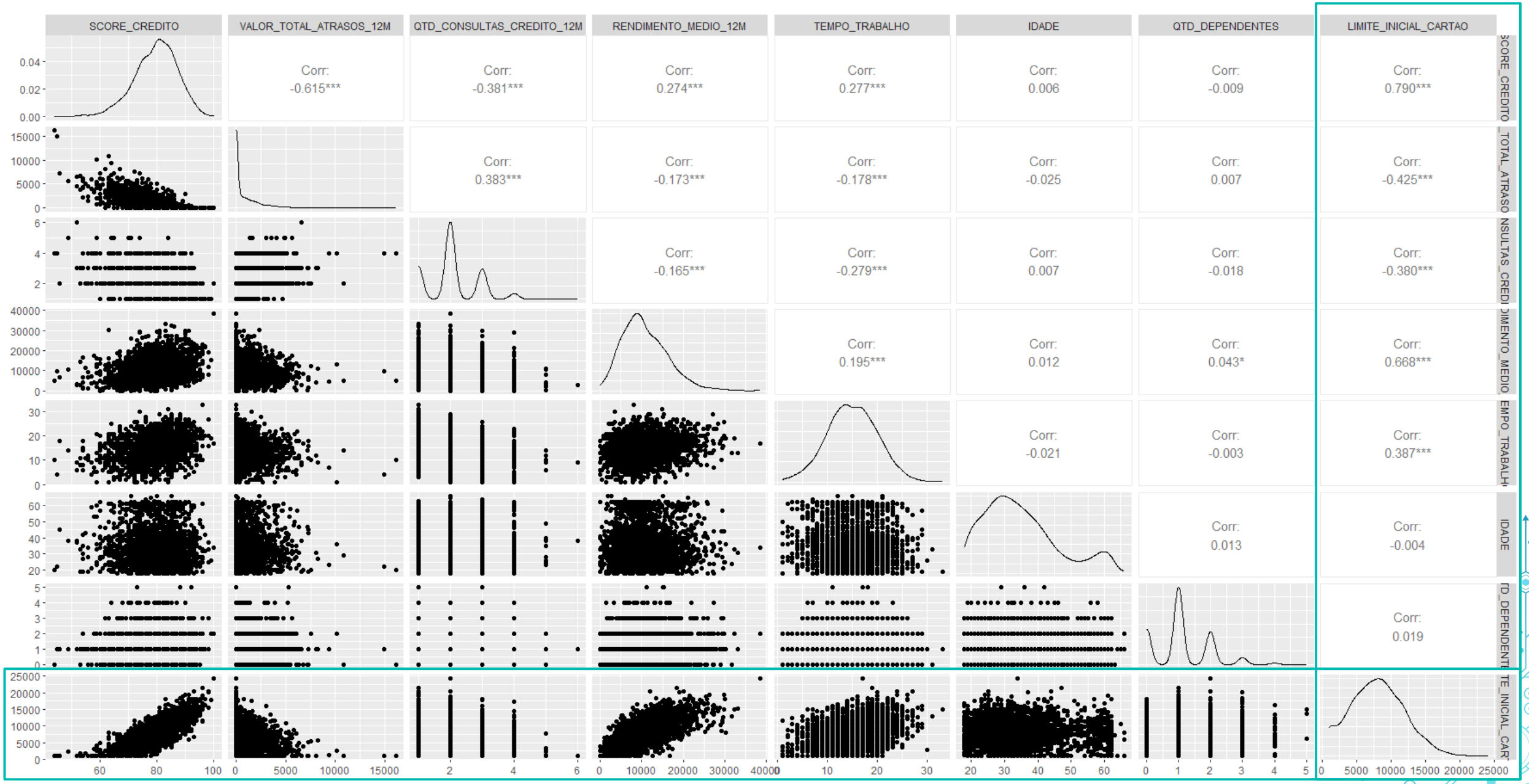
ID_CLIENTE	SCORE_CREDITO	VALOR_TOTAL_ATRASOS_12M	QTD_CONSULTAS_CREDITO_12M	RENDIMENTO_MEDIO_12M	TEMPO_TRABALHO	IDADE	QTD_DEPENDENTES	LIMITE_INICIAL_CARTAO
#0001	79	0	2	9.350	15	37	0	9.700
#0002	73	3.195,52	1	2.010	17	22	0	1.400
#0003	80	0	3	15.660	16	57	2	9.800
#0004	77	0	1	18.640	11	33	0	8.300
#0005	89	0	1	7.550	15	21	1	11.700
#0006	68	4.635,26	3	2.300	9	27	0	1.000
#0007	86	0	1	4.950	9	24	1	8.400
#0008	80	0	1	9.230	18	38	1	4.700
#0009	70	1.750,38	1	7.050	20	43	1	3.000
#0010	81	1.345,54	3	7.780	15	25	0	6.300
...	...	...	...	...	...	...	...	...

Arquivo: Limite_Cartao (.txt)

FIA BUSINESS SCHOOL lab.data

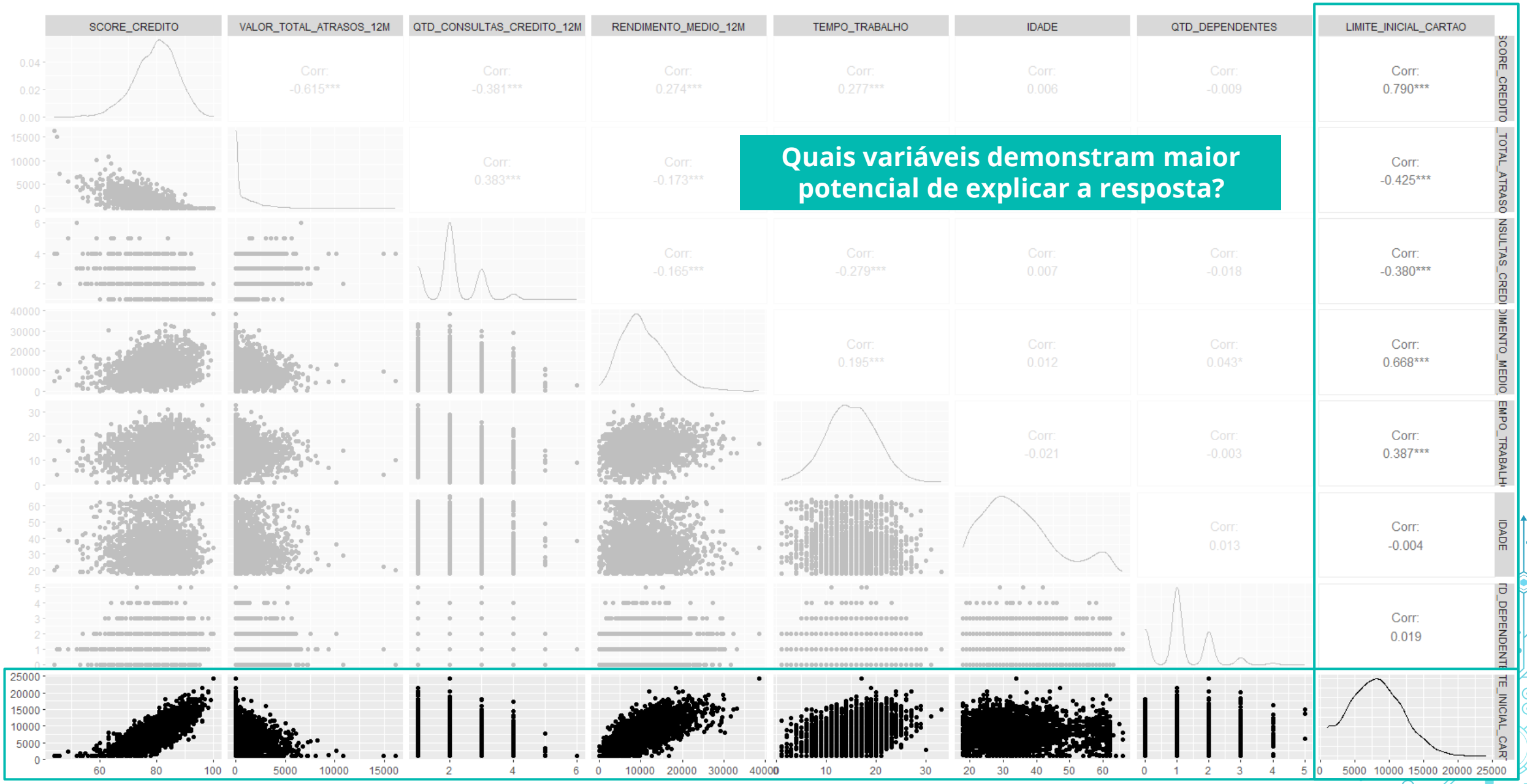
Análise Bidimensional: Correlograma

Variável resposta



Análise Bidimensional: Correlograma

Variável resposta



Case: Limite de Cartão de Crédito

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                           Estimate    Std. Error
(Intercept)              -24152.331107    435.192831
SCORE_CREDITO               348.113952      5.001174
VALOR_TOTAL_ATRASOS_12M       0.319319      0.025934
QTD_CONSULTAS_CREDITO_12M   -275.521607     40.573670
RENDIMENTO_MEDIO_12M          0.335129      0.005469
TEMPO_TRABALHO               86.645754      6.050637
IDADE                        -2.829935      2.526934
QTD_DEPENDENTES              14.081375     33.330662
---


Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```


FIA BUSINESS SCHOOL lab.data

Case: Limite de Cartão de Crédito

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                            Estimate    Std. Error
(Intercept)              -24152.331107    435.192831
SCORE_CREDITO               348.113952      5.001174
VALOR_TOTAL_ATRASOS_12M       0.319319      0.025934
QTD_CONSULTAS_CREDITO_12M   -275.521607     40.573670
RENDIMENTO_MEDIO_12M          0.335129      0.005469
TEMPO_TRABALHO               86.645754      6.050637
IDADE                        -2.829935      2.526934
QTD_DEPENDENTES              14.081375     33.330662
---


Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Algumas medidas resumo dos resíduos

FIA BUSINESS SCHOOL lab.data

Case: Limite de Cartão de Crédito

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                          Estimate   Std. Error
(Intercept)           -24152.331107   435.192831
SCORE_CREDITO            348.113952     5.001174
VALOR_TOTAL_ATRASOS_12M    0.319319     0.025934
QTD_CONSULTAS_CREDITO_12M -275.521607    40.573670
RENDIMENTO_MEDIO_12M       0.335129     0.005469
TEMPO_TRABALHO            86.645754     6.050637
IDADE                     -2.829935     2.526934
QTD_DEPENDENTES           14.081375    33.330662
---


Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Estimativas amostrais $b_0, b_1, \ldots, b_k$ dos respectivos parâmetros populacionais $\beta_0, \beta_1, \ldots, \beta_k$

(neste caso, $k = 7$)

FIA BUSINESS SCHOOL lab.data

Case: Limite de Cartão de Crédito

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                            Estimate   Std. Error
(Intercept)              -24152.331107  435.192831
SCORE_CREDITO               348.113952    5.001174
VALOR_TOTAL_ATRASOS_12M       0.319319    0.025934
QTD_CONSULTAS_CREDITO_12M   -275.521607   40.573670
RENDIMENTO_MEDIO_12M          0.335129    0.005469
TEMPO_TRABALHO               86.645754    6.050637
IDADE                        -2.829935    2.526934
QTD_DEPENDENTES              14.081375   33.330662
---


Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Desvios padrão amostrais de $b_0, b_1, \ldots, b_k$

FIA BUSINESS SCHOOL lab.data

Case: Limite de Cartão de Crédito

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                            Estimate    Std. Error
(Intercept)              -24152.331107    435.192831
SCORE_CREDITO               348.113952      5.001174
VALOR_TOTAL_ATRASOS_12M       0.319319      0.025934
QTD_CONSULTAS_CREDITO_12M   -275.521607     40.573670
RENDIMENTO_MEDIO_12M          0.335129      0.005469
TEMPO_TRABALHO               86.645754      6.050637
IDADE                        -2.829935      2.526934
QTD_DEPENDENTES              14.081375     33.330662
---


Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Coeficiente de determinação (R^2) **ajustado** do modelo

FIA BUSINESS SCHOOL lab.data

Qualidade de Ajuste

Coeficiente de determinação ajustado (R^2)

O coeficiente de determinação (R^2) tradicional sofre de um problema: seu valor **sempre aumenta** quando novas variáveis são adicionadas, saindo da regressão linear simples e passando para a **múltipla**. Isso ocorre mesmo que as variáveis não sejam estatisticamente significativas, ou mesmo que atrapalhem a predição.

Uma alternativa para contornar o problema é considerar o coeficiente $\boldsymbol{R^2}$ **ajustado**, versão corrigida do R^2 para quando temos 2 ou mais variáveis explicativas. Ele pode **decrescer** caso seja acrescentada uma variável que atrapalhe o modelo.

$$R^2\ ajustado = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

Fórmula do R² ajustado

(a letra n corresponde à quantidade de observações, e k à quantidade de variáveis explicativas do modelo)

Processo de Seleção de Variáveis

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)             -24152.331107   435.192831 -55.498 < 0.0000000000000002 ***
SCORE_CREDITO              348.113952     5.001174  69.606 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M      0.319319     0.025934  12.313 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -275.521607    40.573670  -6.791      0.0000000000139 ***
RENDIMENTO_MEDIO_12M          0.335129     0.005469  61.281 < 0.0000000000000002 ***
TEMPO_TRABALHO               86.645754     6.050637  14.320 < 0.0000000000000002 ***
IDADE                        -2.829935     2.526934  -1.120                0.263
QTD_DEPENDENTES              14.081375    33.330662   0.422                0.673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

p-valores dos testes associados a $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_k}$.

Todos os testes avaliam **nulidade** *versus* **não nulidade**:

$$\begin{cases} H\colon \beta_i = 0 \\ A\colon \beta_i \neq 0 \end{cases}$$

para $i = 0, 1, \ldots, k$.

FIA BUSINESS SCHOOL lab.data

Processo de Seleção de Variáveis

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)            -24152.331107   435.192831 -55.498 < 0.0000000000000002 ***
SCORE_CREDITO             348.113952     5.001174  69.606 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M     0.319319     0.025934  12.313 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -275.521607    40.573670  -6.791      0.0000000000139 ***
RENDIMENTO_MEDIO_12M        0.335129     0.005469  61.281 < 0.0000000000000002 ***
TEMPO_TRABALHO             86.645754     6.050637  14.320 < 0.0000000000000002 ***
IDADE                      -2.829935     2.526934  -1.120                0.263
QTD_DEPENDENTES            14.081375    33.330662   0.422                0.673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

As variáveis testadas são **relevantes** para predizer o valor do limite de cartão de crédito, com ao menos 95% de confiança?

FIA BUSINESS SCHOOL lab.data

Processo de Seleção de Variáveis

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)            -24152.331107   435.192831 -55.498 < 0.0000000000000002 ***
SCORE_CREDITO             348.113952     5.001174  69.606 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M     0.319319     0.025934  12.313 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -275.521607    40.573670  -6.791      0.0000000000139 ***
RENDIMENTO_MEDIO_12M        0.335129     0.005469  61.281 < 0.0000000000000002 ***
TEMPO_TRABALHO             86.645754     6.050637  14.320 < 0.0000000000000002 ***
IDADE                      -2.829935     2.526934  -1.120                0.263
QTD_DEPENDENTES            14.081375    33.330662   0.422                0.673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Na regressão linear múltipla, pode acontecer de algumas variáveis serem estatisticamente significativas, e **outras não**.

Portanto, devemos realizar um processo gradual de **redução** de variáveis, até que restem apenas as que forem significativas, de acordo com a confiança desejada. Chamamos este processo de ***stepwise backward.***

Processo de Seleção de Variáveis

Modelo 1: com todas as variáveis explicativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE + QTD_DEPENDENTES, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4858.6  -979.8     1.3   954.0  5107.3

Coefficients:
                          Estimate    Std. Error t value            Pr(>|t|)
(Intercept)           -24152.331107    435.192831 -55.498 < 0.0000000000000002 ***
SCORE_CREDITO            348.113952      5.001174  69.606 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M    0.319319      0.025934  12.313 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -275.521607     40.573670  -6.791      0.0000000000139 ***
RENDIMENTO_MEDIO_12M       0.335129      0.005469  61.281 < 0.0000000000000002 ***
TEMPO_TRABALHO            86.645754      6.050637  14.320 < 0.0000000000000002 ***
IDADE                     -2.829935      2.526934  -1.120                0.263
QTD_DEPENDENTES           14.081375     33.330662   0.422                0.673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2492 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

A **quantidade de dependentes** é a variável menos relevante (maior *p*-valor). Não é possível afirmar que ela está linearmente associada com o limite de cartão de crédito na população, com ao menos 95% de confiança. Logo, ela pode ser **retirada** do modelo.

FIA BUSINESS SCHOOL lab.data

Processo de Seleção de Variáveis

Modelo 2: retirando QTD_DEPENDENTES, que não é estatisticamente significativa

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4860.5  -976.5     6.9   951.1  5092.5

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              -24134.363979   433.038490 -55.733 < 0.0000000000000002 ***
SCORE_CREDITO               348.077713     4.999615  69.621 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M       0.319396     0.025929  12.318 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -275.902025    40.556992  -6.803      0.0000000000128 ***
RENDIMENTO_MEDIO_12M          0.335236     0.005462  61.377 < 0.0000000000000002 ***
TEMPO_TRABALHO               86.618749     6.049302  14.319 < 0.0000000000000002 ***
IDADE                        -2.815910     2.526299  -1.115                0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1411 on 2493 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Processo de Seleção de Variáveis

Modelo 2: retirando QTD_DEPENDENTES, que não é estatisticamente significativa

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4860.5  -976.5     6.9   951.1  5092.5

Coefficients:
                          Estimate    Std. Error t value             Pr(>|t|)
(Intercept)           -24134.363979   433.038490 -55.733 < 0.0000000000000002 ***
SCORE_CREDITO            348.077713     4.999615  69.621 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M    0.319396     0.025929  12.318 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -275.902025    40.556992  -6.803      0.0000000000128 ***
RENDIMENTO_MEDIO_12M       0.335236     0.005462  61.377 < 0.0000000000000002 ***
TEMPO_TRABALHO            86.618749     6.049302  14.319 < 0.0000000000000002 ***
IDADE                     -2.815910     2.526299  -1.115                0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1411 on 2493 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Adotando 95% de confiança, não é possível afirmar que a **idade do cliente** está linearmente associada com o limite de cartão de crédito na população. Logo, ela também pode ser **retirada** do modelo.

Processo de Seleção de Variáveis

Modelo 2: retirando QTD_DEPENDENTES, que não é estatisticamente significativa

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO +
    IDADE, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4860.5  -976.5     6.9   951.1  5092.5

Coefficients:
                          Estimate    Std. Error t value             Pr(>|t|)
(Intercept)           -24134.363979   433.038490 -55.733 < 0.0000000000000002 ***
SCORE_CREDITO            348.077713     4.999615  69.621 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M    0.319396     0.025929  12.318 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -275.902025    40.556992  -6.803      0.0000000000128 ***
RENDIMENTO_MEDIO_12M       0.335236     0.005462  61.377 < 0.0000000000000002 ***
TEMPO_TRABALHO            86.618749     6.049302  14.319 < 0.0000000000000002 ***
IDADE                     -2.815910     2.526299  -1.115                0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1411 on 2493 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8664
```

Não poderíamos ter retirado **quantidade de dependentes** e **idade** ambas no primeiro passo?

Não, devemos retirar apenas **1 variável por vez**, do maior para o menor *p*-valor, pois a retirada de uma variável pode alterar o grau de influência associado às demais.

FIA BUSINESS SCHOOL lab.data

Processo de Seleção de Variáveis

Modelo 3: retirando QTD_DEPENDENTES e IDADE, que não são estatisticamente significativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO,
    data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4841.2  -975.5    10.5   950.3  5095.7

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              -24236.386302   423.275434 -57.259 < 0.0000000000000002 ***
SCORE_CREDITO               348.118442     4.999724  69.628 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M       0.320249     0.025919  12.356 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -276.431359    40.556183  -6.816      0.0000000000117 ***
RENDIMENTO_MEDIO_12M          0.335149     0.005462  61.364 < 0.0000000000000002 ***
TEMPO_TRABALHO               86.771740     6.048039  14.347 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2494 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.8664
```


FIA BUSINESS SCHOOL lab.data

Processo de Seleção de Variáveis

Modelo 3: retirando QTD_DEPENDENTES e IDADE, que não são estatisticamente significativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO,
    data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4841.2  -975.5    10.5   950.3  5095.7

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              -24236.386302   423.275434 -57.259 < 0.0000000000000002 ***
SCORE_CREDITO               348.118442     4.999724  69.628 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M       0.320249     0.025919  12.356 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -276.431359    40.556183  -6.816      0.0000000000117 ***
RENDIMENTO_MEDIO_12M          0.335149     0.005462  61.364 < 0.0000000000000002 ***
TEMPO_TRABALHO               86.771740     6.048039  14.347 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2494 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.8664
```

Agora, **todas as variáveis restantes** são estatisticamente significativas no modelo, com ao menos 95% de confiança.

Note que não houve alteração no R^2 ajustado, que é de cerca de **87%**.

Processo de Seleção de Variáveis

Modelo 3: retirando QTD_DEPENDENTES e IDADE, que não são estatisticamente significativas

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + VALOR_TOTAL_ATRASOS_12M +
    QTD_CONSULTAS_CREDITO_12M + RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO,
    data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4841.2  -975.5    10.5   950.3  5095.7

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              -24236.386302   423.275434 -57.259 < 0.0000000000000002 ***
SCORE_CREDITO               348.118442     4.999724  69.628 < 0.0000000000000002 ***
VALOR_TOTAL_ATRASOS_12M       0.320249     0.025919  12.356 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -276.431359    40.556183  -6.816      0.0000000000117 ***
RENDIMENTO_MEDIO_12M          0.335149     0.005462  61.364 < 0.0000000000000002 ***
TEMPO_TRABALHO               86.771740     6.048039  14.347 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 2494 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.8664
```

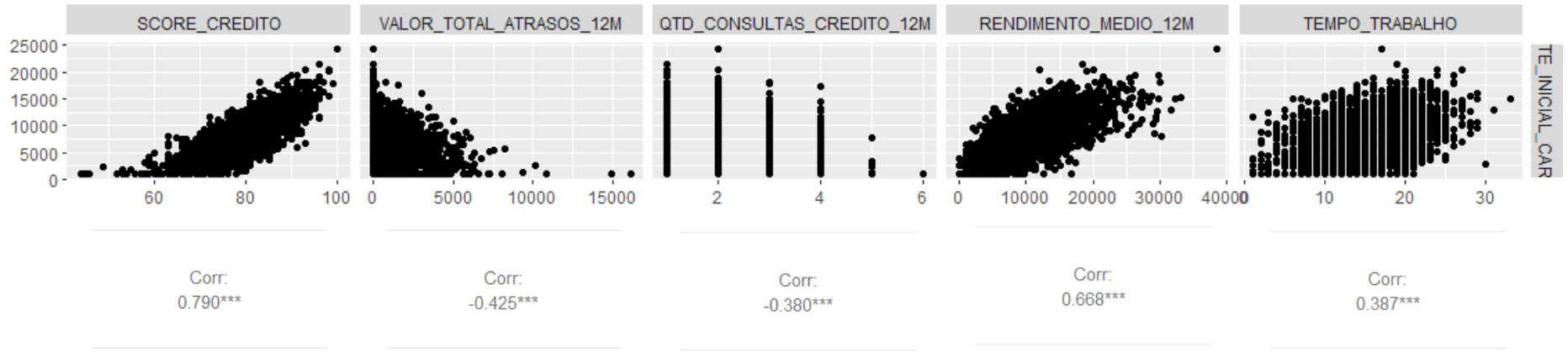
Por fim, os parâmetros estimados são coerentes com a **visão de negócio** e com o que havíamos concluído na **análise bidimensional**?

Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                                Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

Análise bivariada:



FIA BUSINESS SCHOOL lab.data

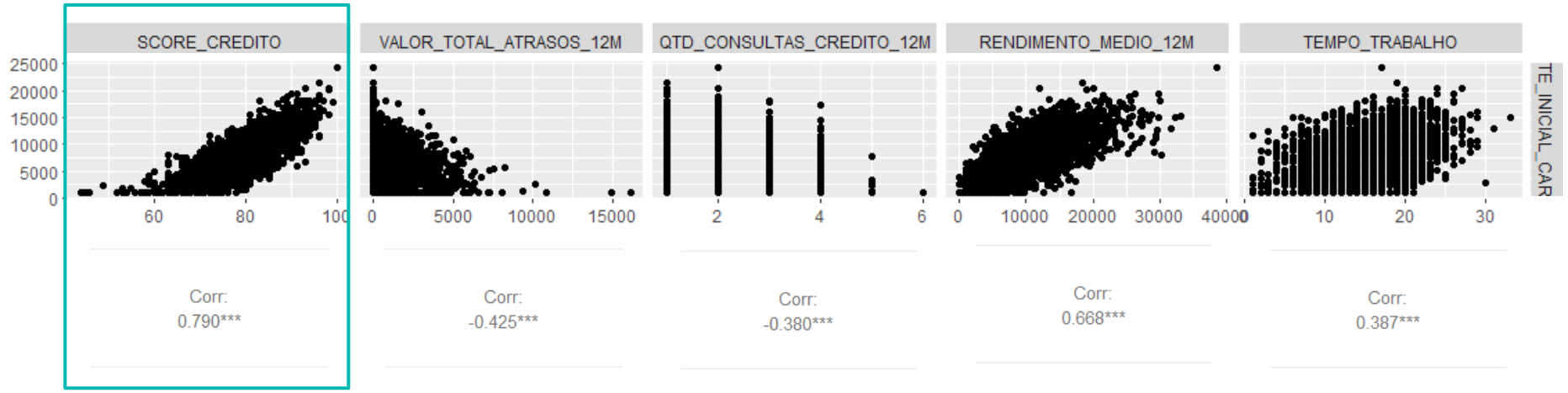
Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                                Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

- A cada 1 ponto de ***score* de crédito** a mais que um cliente tenha, seu limite de cartão de crédito **aumenta**, em média, **R$ 348,12**, se mantidas fixas as demais características.
- Essa interpretação é **coerente** com a análise bivariada, na qual observamos uma correlação linear positiva (r = 0,790).

Análise bivariada:



Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                                Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

- A cada 1 **consulta** a mais no nome de um cliente no último ano, seu limite de cartão de crédito **diminui**, em média, **R$ 276,43**, se mantidas fixas as demais características.
- Essa interpretação é **coerente** com a análise bivariada, na qual observamos uma correlação linear negativa (r = -0,380).

Análise bivariada:

FIA BUSINESS SCHOOL lab.data

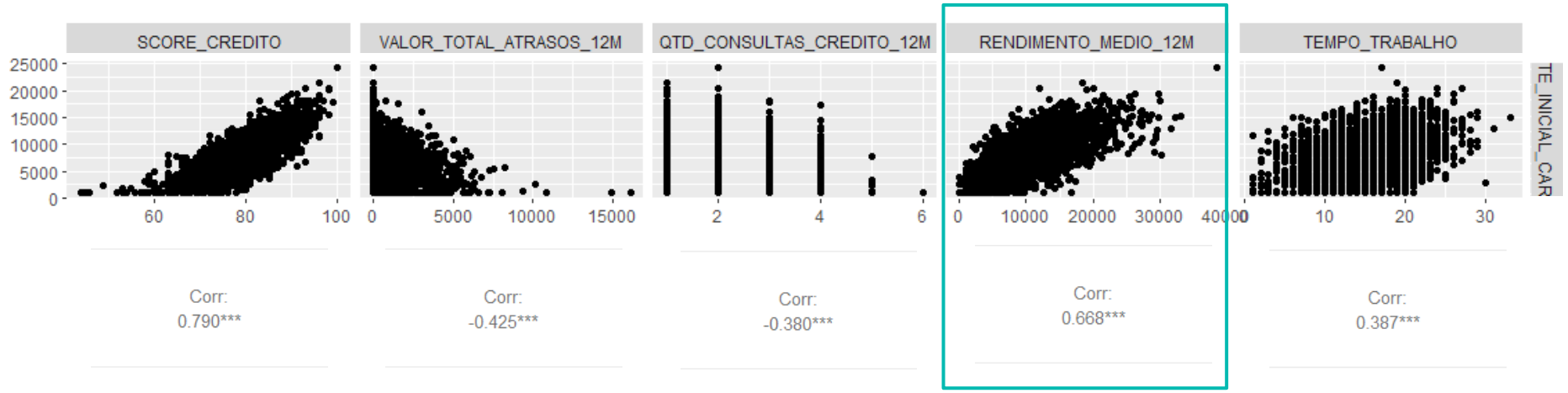
Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                              Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

- A cada 1 real a mais de **rendimento médio mensal** de um cliente, seu limite de cartão de crédito **aumenta**, em média, **R$ 0,34**, se mantidas fixas as demais características.
- Essa interpretação é **coerente** com a análise bivariada, na qual observamos uma correlação linear positiva (r = 0,668).

Análise bivariada:



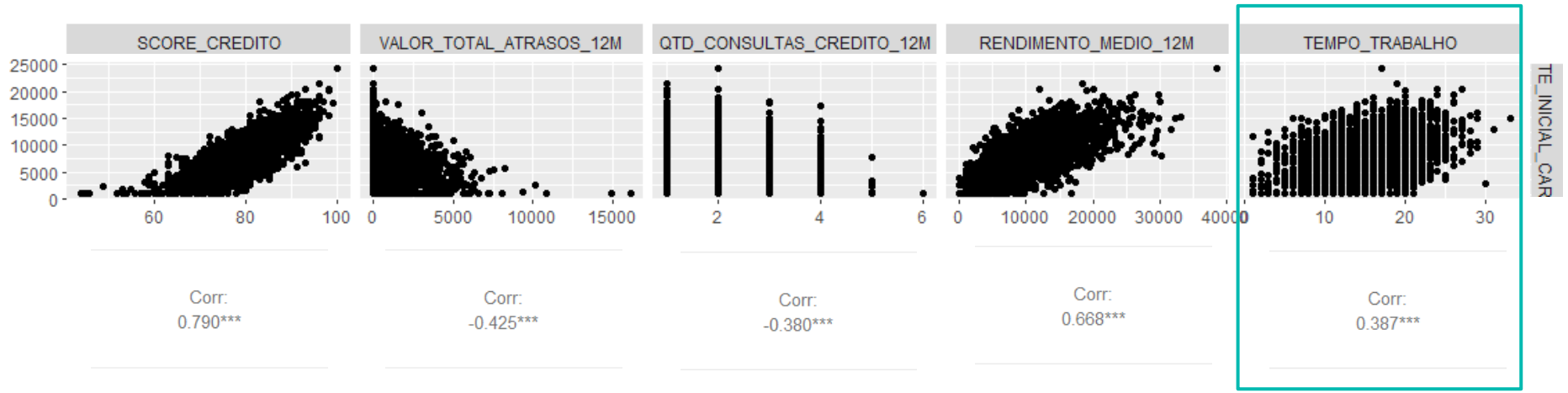
Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                                Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

- A cada 1 ano a mais de **tempo de trabalho** que um cliente tenha, seu limite de cartão de crédito **aumenta**, em média, **R$ 86,77**, se mantidas fixas as demais características.
- Essa interpretação é **coerente** com a análise bivariada, na qual observamos uma correlação linear positiva (r = 0,387).

Análise bivariada:



Interpretação de Parâmetros Estimados

Intepretação dos parâmetros estimados no modelo 3

```
Coefficients:
                                Estimate
(Intercept)                -24236.386302
SCORE_CREDITO                348.118442
VALOR_TOTAL_ATRASOS_12M        0.320249
QTD_CONSULTAS_CREDITO_12M   -276.431359
RENDIMENTO_MEDIO_12M           0.335149
TEMPO_TRABALHO                86.771740
```

- A cada 1 real a mais em **atrasos** de um cliente nos últimos 12m, seu limite de cartão de crédito **aumenta**, em média, **R$ 0,32**, se mantidas fixas as demais características.
- Essa interpretação **não é coerente** com a análise bivariada, na qual observamos uma correlação linear negativa (r = -0,425).

  O que houve aqui?

Análise bivariada:

FIA BUSINESS SCHOOL lab.data

Análise Bidimensional: Correlograma

As variáveis referentes a **atrasos** e ***score* de crédito** possuem correlação moderada a alta entre si. Isso pode ser um problema?

Colinearidade

A **colinearidade** ocorre quando existe **alta correlação linear** entre duas **variáveis explicativas** no modelo. É também chamada de multicolinearidade, quando envolve mais de duas variáveis.

Devido a este problema:

- Não é possível determinar o coeficiente β (“peso”) ideal para cada variável individualmente, dado que possuem alta **redundância** entre si.
- O processo de estimação se torna **instável**, sendo comum observar estimativas com valores **atípicos** (extremamente altos ou baixos) e/ou com **sinais trocados** em relação ao esperado pelo contexto e pela análise bidimensional inicial.

Em geral, é possível retirar as variáveis envolvidas no problema de colinearidade sem perda relevante no $\boldsymbol{R^2}$ **ajustado** e sem prejuízo para a **interpretação** das demais variáveis do modelo.

Case: Limite de Cartão de Crédito

Modelo 4: retirando também VALOR_TOTAL_ATRASOS_12M, devido a problema de colinearidade

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + QTD_CONSULTAS_CREDITO_12M +
    RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4779.9  -994.1   -17.8   987.4  6871.9

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              -21575.465956   375.307208 -57.487 < 0.0000000000000002 ***
SCORE_CREDITO               314.863503     4.339666  72.555 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M   -173.482292    40.879674  -4.244            0.0000228 ***
RENDIMENTO_MEDIO_12M           0.335432     0.005625  59.631 < 0.0000000000000002 ***
TEMPO_TRABALHO                88.984299     6.226423  14.291 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1454 on 2495 degrees of freedom
Multiple R-squared:  0.8585,    Adjusted R-squared:  0.8583
```

Arquivo: Limite_Cartao (.txt)

Case: Limite de Cartão de Crédito

Modelo 4: retirando também VALOR_TOTAL_ATRASOS_12M, devido a problema de colinearidade

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + QTD_CONSULTAS_CREDITO_12M +
    RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4779.9  -994.1   -17.8   987.4  6871.9

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)            -21575.465956   375.307208 -57.487 < 0.0000000000000002 ***
SCORE_CREDITO              314.863503     4.339666  72.555 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -173.482292    40.879674  -4.244            0.0000228 ***
RENDIMENTO_MEDIO_12M         0.335432     0.005625  59.631 < 0.0000000000000002 ***
TEMPO_TRABALHO              88.984299     6.226423  14.291 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1454 on 2495 degrees of freedom
Multiple R-squared:  0.8585,    Adjusted R-squared:  0.8583
```

Redução ínfima de R^2, de **86,6%** para **85,8%**

Arquivo: Limite_Cartao (.txt)

FIA BUSINESS SCHOOL lab.data

Estatística VIF

A estatística **VIF** (*variance inflation factor*, ou fator de inflação de variância) é uma alternativa para mensurar a presença de **colinearidade** no modelo de regressão.

O valor do VIF para cada variável explicativa X_i é dado por:

$$VIF_i = \frac{1}{1 - R_i^2}$$

onde R_i^2 representa o coeficiente de determinação de um modelo de regressão que considera X_i como **variável resposta**, e as demais variáveis como explicativas (exceto a variável resposta original Y).

O valor do VIF varia de **1** a **mais infinito**. A partir de estudos anteriores, valores de VIF **próximos ou maiores que 2** já revelam sinais de colinearidade (ou seja, quando $R_i^2 > 0{,}50$).

Estatística VIF

A estatística **VIF** (*variance inflation factor*, ou fator de inflação de variância) é uma alternativa para mensurar a presença de **colinearidade** no modelo de regressão.

No *case* de limite de cartão de crédito:

```
vif(modelo_3)
```

SCORE_CREDITO	VALOR_TOTAL_ATRASOS_12M	QTD_CONSULTAS_CREDITO_12M	RENDIMENTO_MEDIO_12M	TEMPO_TRABALHO
1.786745	1.680847	1.272404	1.101425	1.142733

No **modelo 3**, obtemos valores de VIF mais próximos de 2 justamente para as variáveis explicativas com alto grau de correlação entre si: ***score* de crédito** e **valor total de atrasos em 12 meses**.

```
vif(modelo_4)
```

SCORE_CREDITO	QTD_CONSULTAS_CREDITO_12M	RENDIMENTO_MEDIO_12M	TEMPO_TRABALHO
1.268978	1.218700	1.101406	1.141732

Ao retirar esta última variável no **modelo 4**, houve redução do VIF associado à variável ***score* de crédito**.

Interpretação de Parâmetros Estimados

Equação e interpretação do **modelo final** (modelo 4)

$$\hat{y} = -21.575{,}47 + 314{,}86 \cdot x_1 - 173{,}48 \cdot x_2 + 0{,}34 \cdot x_3 + 88{,}98 \cdot x_4$$

Sendo que:

- **-R$ 21.575,47** é o limite de cartão de crédito "basal", sem interpretação prática, correspondente a um cliente que tivesse valor zero em todas as variáveis explicativas
- **R$ 314,86** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 ponto no *score* de crédito do cliente (mantendo fixas as demais variáveis)
- **R$ 173,48** é a diminuição média no limite de cartão de crédito proporcionado por cada aumento de 1 consulta de crédito realizada no nome do cliente (mantendo fixas as demais variáveis)
- **R$ 0,34** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 real no rendimento médio mensal do cliente (mantendo fixas as demais variáveis)
- **R$ 88,98** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 ano no tempo de trabalho do cliente (mantendo fixas as demais variáveis)
- O ***R²*** indica que **85,8%** do comportamento do limite de cartão de crédito é devidamente capturado/explicado por meio das 4 variáveis explicativas do modelo

FIA BUSINESS SCHOOL lab.data

Ranking de Importância das Variáveis

Equação e interpretação do **modelo final** (modelo 4)

$$\hat{y} = -21.575{,}47 + 314{,}86 \cdot x_1 - 173{,}48 \cdot x_2 + 0{,}34 \cdot x_3 + 88{,}98 \cdot x_4$$

Sendo que:

- **-R$ 21.575,47** é o limite de cartão de crédito "basal", sem interpretação prática, correspondente a um cliente que tivesse valor zero em todas as variáveis explicativas
- **R$ 314,86** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 ponto no *score* de crédito do cliente (mantendo fixas as demais variáveis)
- **R$ 173,48** é a diminuição média no limite de cartão de crédito proporcionado por cada aumento de 1 consulta de crédito realizada no nome do cliente (mantendo fixas as demais variáveis)
- **R$ 0,34** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 real no rendimento médio mensal do cliente (mantendo fixas as demais variáveis)
- **R$ 88,98** é o aumento médio no limite de cartão de crédito proporcionado por cada aumento de 1 ano no tempo de trabalho do cliente (mantendo fixas as demais variáveis)
- O ***R²*** indica que **85,8%** do comportamento do limite de cartão de crédito é devidamente capturado/explicado por meio das 4 variáveis explicativas do modelo

Quais variáveis são as mais importantes no modelo?

FIA BUSINESS SCHOOL lab.data

Ranking de Importância das Variáveis

Modelo 4

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + QTD_CONSULTAS_CREDITO_12M +
    RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4779.9  -994.1   -17.8   987.4  6871.9

Coefficients:
                           Estimate    Std. Error t value             Pr(>|t|)
(Intercept)            -21575.465956   375.307208 -57.487 < 0.0000000000000002 ***
SCORE_CREDITO              314.863503     4.339666  72.555 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M  -173.482292    40.879674  -4.244            0.0000228 ***
RENDIMENTO_MEDIO_12M          0.335432     0.005625  59.631 < 0.0000000000000002 ***
TEMPO_TRABALHO               88.984299     6.226423  14.291 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1454 on 2495 degrees of freedom
Multiple R-squared:  0.8585,    Adjusted R-squared:  0.8583
```

A importância das variáveis deve ser julgada a partir do ***p*-valor**. Quanto menor, mais importante é a variável.

Caso haja empate de *p*-valores, podemos considerar o ***t value***. Quanto maior o seu valor absoluto, mais importante é a variável.

FIA BUSINESS SCHOOL lab.data

Ranking de Importância das Variáveis

Modelo 4

```
Call:
lm(formula = LIMITE_INICIAL_CARTAO ~ SCORE_CREDITO + QTD_CONSULTAS_CREDITO_12M +
    RENDIMENTO_MEDIO_12M + TEMPO_TRABALHO, data = dados_limite)

Residuals:
    Min      1Q  Median      3Q     Max
-4779.9  -994.1   -17.8   987.4  6871.9

Coefficients:
                          Estimate    Std. Error t value             Pr(>|t|)
(Intercept)           -21575.465956    375.307208 -57.487 < 0.0000000000000002 ***
SCORE_CREDITO             314.863503      4.339666  72.555 < 0.0000000000000002 ***
QTD_CONSULTAS_CREDITO_12M -173.482292     40.879674  -4.244            0.0000228 ***
RENDIMENTO_MEDIO_12M         0.335432      0.005625  59.631 < 0.0000000000000002 ***
TEMPO_TRABALHO              88.984299      6.226423  14.291 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1454 on 2495 degrees of freedom
Multiple R-squared:  0.8585,    Adjusted R-squared:  0.8583
```

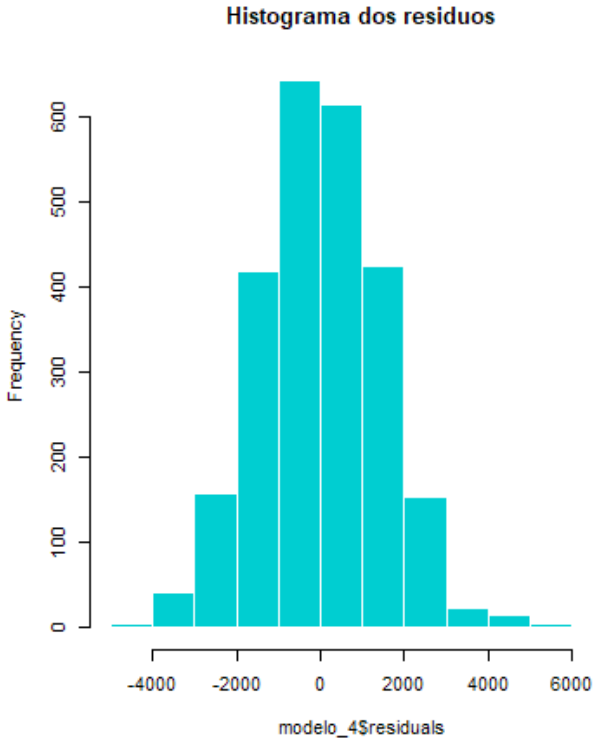
Ranking de importância das variáveis

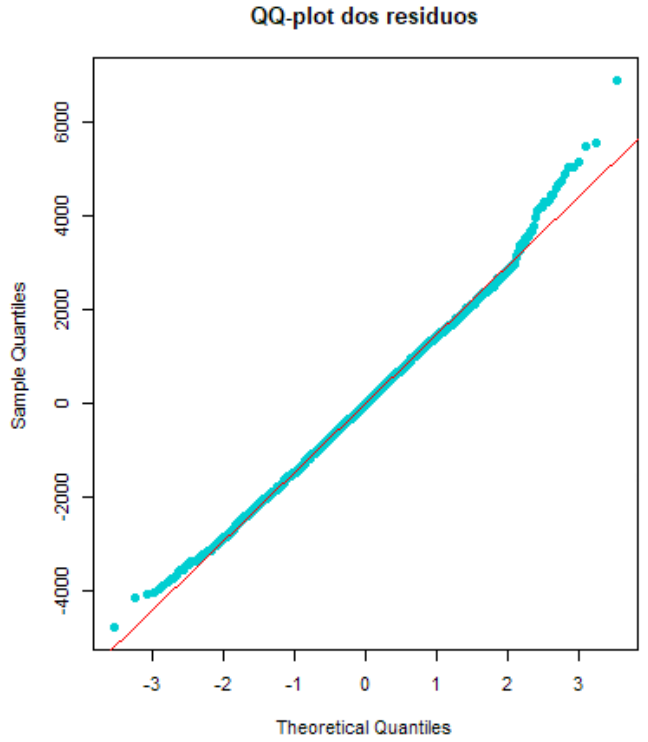
- SCORE_CREDITO → 1ª
- QTD_CONSULTAS_CREDITO_12M → 4ª
- RENDIMENTO_MEDIO_12M → 2ª
- TEMPO_TRABALHO → 3ª

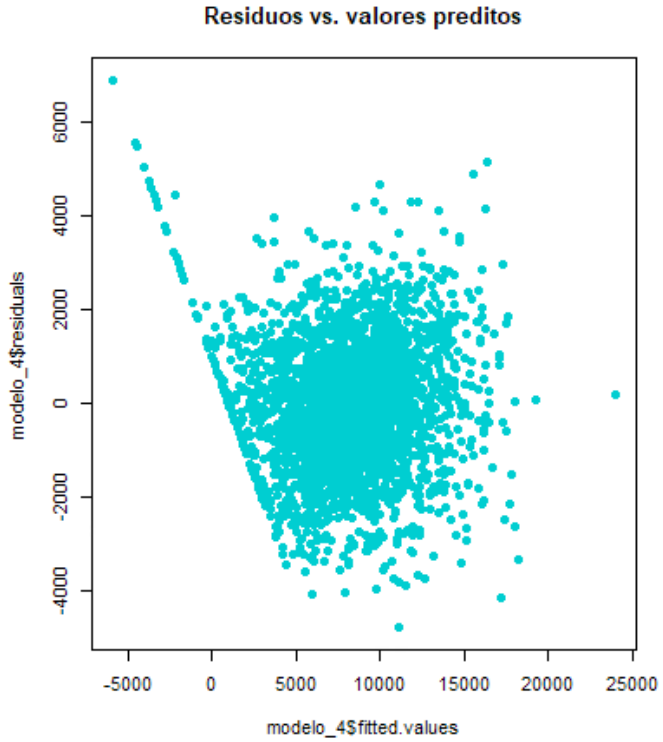
FIA BUSINESS SCHOOL lab.data

Análise de Resíduos

Modelo 4







Os resíduos seguem uma distribuição aproximadamente **simétrica** em torno do zero (dois gráficos da esquerda), e apresentam **variabilidade constante** *versus* os valores preditos (gráfico à direita)
Ponto de atenção: alguns clientes têm limite de cartão de crédito predito negativo.

FIA BUSINESS SCHOOL lab.data

Case: Salário de Funcionários

O time de recrutamento de uma empresa deseja compreender se o **salário inicial** definido pelos gestores para os analistas júniores está associado a determinadas características dos candidatos durante o processo seletivo. Para isso, examinou-se a base de dados dos funcionários contratados nos últimos 12 meses para cargos júniores.

Variável	Descrição
ID_FUNCIONARIO	Código identificador do funcionário que foi contratado como analista júnior
PRETENSAO_SALARIAL	Pretensão salarial (R$) informada no processo seletivo
DESEMPENHO_PROVA	Desempenho do funcionário na prova de admissão: razoável; bom; ou excelente
FLAG_EXPERIENCIA	Indicadora de se o candidato possui experiência anterior na área (1 se sim; 0 se não)
SALARIO_DEFINIDO	Salário (R$) inicial definido na contratação do funcionário

Arquivo: Salario_Funcionarios (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Salário de Funcionários

O time de recrutamento de uma empresa deseja compreender se o **salário inicial** definido pelos gestores para os analistas júniores está associado a determinadas características dos candidatos durante o processo seletivo. Para isso, examinou-se a base de dados dos funcionários contratados nos últimos 12 meses para cargos júniores.

ID_FUNCIONARIO	PRETENSAO_ SALARIAL	DESEMPENHO_ PROVA	FLAG_ EXPERIENCIA	SALARIO_ DEFINIDO
1	4.000	Razoável	Não	3.300
2	3.800	Ótimo	Sim	4.300
3	4.000	Ótimo	Não	3.800
4	5.000	Razoável	Não	4.300
5	2.500	Bom	Não	3.200
6	3.000	Ótimo	Não	4.000
7	4.500	Razoável	Não	3.800
8	2.500	Bom	Não	3.000
9	5.000	Ótimo	Sim	5.000
10	4.000	Bom	Sim	4.300
...	...	...	...	...

Arquivo: Salario_Funcionarios (.txt)

FIA BUSINESS SCHOOL lab.data

Incorporando Variáveis Qualitativas

Agora, vamos incorporar variáveis explicativas **qualitativas** na regressão linear múltipla. Elas devem ser transformadas em variáveis ***dummies*** binárias, que assumem os valores **0 ou 1** a depender da presença de cada característica.

No *case* de salários de funcionários, temos as seguintes variáveis qualitativas:

✓ Variável **FLAG_EXPERIENCIA**, com duas categorias:

- Não
- Sim

→ Cria-se **uma variável *dummy***, que recebe:

- valor **0 (zero)** para funcionários sem experiência
- valor **1 (um)** para funcionários com experiência

✓ Variável **DESEMPENHO_PROVA**, com três categorias:

- Razoável
- Bom
- Ótimo

→ Cria-se **duas variáveis *dummy***:

Uma para a categoria **"bom"**, que recebe **1** caso o funcionário tenha desempenho bom, e **0** caso contrário.

Uma para a categoria **"ótimo"**, que recebe **1** caso o funcionário tenha desempenho ótimo, e **0** caso contrário.

Incorporando Variáveis Qualitativas

Agora, vamos incorporar variáveis explicativas **qualitativas** na regressão linear múltipla. Elas devem ser transformadas em variáveis ***dummies*** binárias, que assumem os valores **0 ou 1** a depender da presença de cada característica.

No *case* de salários de funcionários, temos as seguintes variáveis qualitativas:

✓ Variável **FLAG_EXPERIENCIA**, com duas categorias:

- Não
- Sim

➢ Por que basta criar **1** em vez de **2** *dummies* para FLAG_EXPERIENCIA?

✓ Variável **DESEMPENHO_PROVA**, com três categorias:

- Razoável
- Bom
- Ótimo

➢ Por que basta criar **1** em vez de **3** *dummies* para DESEMPENHO_PROVA?

Incorporando Variáveis Qualitativas

Agora, vamos incorporar variáveis explicativas **qualitativas** na regressão linear múltipla. Elas devem ser transformadas em variáveis ***dummies*** binárias, que assumem os valores **0 ou 1** a depender da presença de cada característica.

Observações:

- ✓ Note que a quantidade de *dummies* necessárias para representar uma variável qualitativa é sempre igual à **quantidade de categorias da variável original menos 1 (um)**.

- ✓ Caso não haja uma ordenação explícita das categorias da variável no R, será assumida como **referência** a primeira categoria **em ordem alfabética**. Para ela, não será criada uma variável *dummy.* Esta categoria é representada pela situação em que todas as demais *dummies* assumem o valor **0 (zero)**.

- ✓ Já se a variável qualitativa for **ordinal** e suas categorias forem ordenadas explicitamente no R, a primeira categoria da ordenação será assumida como referência.

  Obs.: Isso foi feito para a variável DESEMPENHO_PROVA no código em R disponibilizado, por meio da função *factor().*

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)
(Intercept)       1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL   0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom 254.02894  100.16123   2.536              0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim 371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

Variáveis qualitativas

Note que todas as variáveis são **estatisticamente significativas** no modelo, com ao menos 95% de confiança. Então, não é necessário reduzir o modelo.

Arquivo: Salario_Funcionarios (.txt)

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL      0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom   254.02894  100.16123   2.536               0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim   371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

O intercepto indica que ~**R$ 1.901** é o salário médio (hipotético) para um analista com pretensão salarial zero, desempenho razoável na prova de seleção e sem experiência anterior na área.

Arquivo: Salario_Funcionarios (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)         1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL     0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom  254.02894  100.16123   2.536               0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim  371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

A cada aumento de 1 real na **pretensão salarial**, o salário do funcionário aumenta ~**R$ 0,42**, em média, mantidas fixas as demais variáveis do modelo.

Arquivo: Salario_Funcionarios (.txt)

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL      0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom   254.02894  100.16123   2.536               0.0158 *
DESEMPENHO_PROVAOtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim   371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

Quando o **desempenho na prova** passa de *razoável* (referência) para *bom*, o salário do funcionário aumenta ~**R$ 254**, em média, mantidas fixas as demais variáveis do modelo.

Arquivo: Salario_Funcionarios (.txt)

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)
(Intercept)          1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL      0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom   254.02894  100.16123   2.536              0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim   371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

Quando o **desempenho na prova** passa de *razoável* (referência) para *ótimo*, o salário do funcionário aumenta ~**R$ 514**, em média, mantidas fixas as demais variáveis do modelo.

Arquivo: Salario_Funcionarios (.txt)

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL      0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom   254.02894  100.16123   2.536               0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim   371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

Quando a **experiência anterior** na área passa de *não* (referência) para *sim*, o salário do funcionário aumenta ~**R$ 371**, em média, mantidas fixas as demais variáveis do modelo.

Arquivo: Salario_Funcionarios (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Salário de Funcionários

Modelo

```
Call:
lm(formula = SALARIO_DEFINIDO ~ PRETENSAO_SALARIAL + DESEMPENHO_PROVA +
    FLAG_EXPERIENCIA, data = dados_salario)

Residuals:
    Min      1Q  Median      3Q     Max
-284.33 -176.48   -6.06  133.87  349.55

Coefficients:
                     Estimate Std. Error t value          Pr(>|t|)
(Intercept)          1900.66926  189.03543  10.055 0.00000000000735 ***
PRETENSAO_SALARIAL      0.41734    0.04054  10.295 0.00000000000394 ***
DESEMPENHO_PROVABom   254.02894  100.16123   2.536            0.0158 *
DESEMPENHO_PROVAÓtimo 514.30140   90.95822   5.654 0.00000220559499 ***
FLAG_EXPERIENCIASim   371.48636   67.49235   5.504 0.00000348033385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.3 on 35 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7759
```

77,6% de variabilidade explicada pelo modelo

Arquivo: Salario_Funcionarios (.txt)

Case: Salário de Funcionários

O time de recrutamento de uma empresa deseja compreender se o **salário inicial** definido pelos gestores para os analistas júniores está associado a características manifestadas pelos candidatos durante o processo seletivo. Para isso, examinou-se a base de dados dos funcionários contratados nos últimos 12 meses para o cargo.

a) Faça uma breve análise exploratória da base de dados.
b) De forma gráfica, parece existir relação linear entre o salário inicial dos analistas e as demais variáveis? Calcule e interprete o coeficiente de correlação linear entre o salário inicial e a pretensão salarial.
c) Construa um modelo de regressão linear múltipla e faça a seleção de variáveis pelo método *stepwise backward*, considerando 95% de confiança. Atente-se à colinearidade por meio do índice VIF, e remova as variáveis envolvidas, se necessário.
d) Interprete as estimativas dos parâmetros, os intervalos de 95% de confiança e os *p*-valores.
e) Escreva a equação estimada do modelo final.
f) Interprete o valor do coeficiente de determinação ajustado (R^2). Como você avalia a qualidade do modelo?
g) Analise graficamente os resíduos do modelo. Eles seguem um comportamento razoável?
h) Estime o salário inicial para um funcionário com pretensão salarial de R$ 3.000, com ótimo desempenho na prova de seleção, e com experiência anterior na área.

Arquivo: Salario_Funcionarios (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Satisfação em Restaurante

Uma rede de restaurantes deseja investigar se a **quantidade de clientes** recebidos em cada unidade no mês de dezembro está associada a aspectos gerais de atendimento e de satisfação dos clientes, demonstrados ao longo dos meses de janeiro a novembro do mesmo ano. Para isso, organizou uma base de dados com diversas informações, para cada uma de suas 60 unidades.

Variável	Descrição
ID_UNIDADE	Identificação da unidade
FLAG_LITORAL	Indica se a unidade está situada em litoral, sim ou não
TEMPO_MED_ESPERA	Tempo médio (em min) de espera do cliente, entre o pedido e a chegada do prato principal, de jan a nov
VALOR_MED_CONTA	Valor médio (R$) das contas dos clientes
NOTA_MED_COMIDA	Nota média (de 0 a 10) de satisfação dos clientes com a qualidade da comida, de jan a nov
NOTA_MED_ATENDIMENTO	Nota média (de 0 a 10) de satisfação dos clientes com a qualidade do atendimento, de jan a nov
NOTA_MED_AMBIENTE	Nota média (de 0 a 10) de satisfação dos clientes com a qualidade do ambiente, de jan a nov
QTDE_MED_CLIENTES	Quantidade média mensal de clientes recebidos, de jan a nov
QTDE_CLIENTES_DEZ	Quantidade de clientes recebidos em dezembro

Arquivo: Satisfacao_Restaurante (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Satisfação em Restaurante

Uma rede de restaurantes deseja investigar se a **quantidade de clientes** recebidos em cada unidade no mês de dezembro está associada a aspectos gerais de atendimento e de satisfação dos clientes, demonstrados ao longo dos meses de janeiro a novembro do mesmo ano. Para isso, organizou uma base de dados com diversas informações, para cada uma de suas 60 unidades.

ID_UNIDADE	FLAG_ LITORAL	TEMPO_MED_ ESPERA	VALOR_MED_ CONTA	NOTA_MED_ COMIDA	NOTA_MED_ ATENDIMENTO	NOTA_MED_ AMBIENTE	QTDE_MED_ CLIENTES	QTDE_ CLIENTES_DEZ
Unidade Aeroporto (SP)	0	17	131,10	9	7	10	2.960	3.470
Unidade Água Branca (SP)	0	19	189,30	9	7	10	3.200	2.700
Unidade Alto de Pinheiros (SP)	0	10	192,00	9	10	10	3.640	9.250
Unidade Barra da Tijuca (RJ)	1	10	256,10	9	8	9	3.760	12.210
Unidade Bela Vista (SP)	0	16	174,00	9	9	9	6.200	9.040
Unidade Belvedere (BH)	0	12	211,40	9	7	6	3.500	1.920
Unidade Bertioga (SP)	1	21	210,40	9	9	9	1.040	2.830
Unidade Botafogo (RJ)	1	16	174,00	9	10	9	4.180	14.240
Unidade Brooklin (SP)	0	14	166,00	10	6	10	4.640	9.010
Unidade Butantã (SP)	0	15	157,50	8	9	9	1.820	5.910
...	...	...	...	...	...	...	...	...

Arquivo: Satisfacao_Restaurante (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Satisfação em Restaurante

Uma rede de restaurantes deseja investigar se a **quantidade de clientes** recebidos em cada unidade no mês de dezembro está associada a aspectos gerais de atendimento e de satisfação dos clientes, demonstrados ao longo dos meses de janeiro a novembro do mesmo ano. Para isso, organizou uma base de dados com diversas informações, para cada uma de suas 60 unidades.

a) Faça uma breve análise exploratória da base de dados.
b) De forma gráfica, parece existir relação linear entre a quantidade de clientes em dezembro e as demais variáveis? Calcule e interprete os coeficientes de correlação linear entre os pares de variáveis quantitativas.
c) Construa um modelo de regressão linear múltipla e faça a seleção de variáveis pelo método *stepwise backward*, considerando 95% de confiança. Atente-se à colinearidade por meio do índice VIF, e remova as variáveis envolvidas, se necessário.
d) Interprete as estimativas dos parâmetros, os intervalos de 95% de confiança e os *p*-valores.
e) Escreva a equação estimada do modelo final.
f) Interprete o valor do coeficiente de determinação ajustado (R^2). Como você avalia a qualidade do modelo?
g) Analise graficamente os resíduos do modelo. Eles seguem um comportamento razoável?
h) Estime a quantidade de clientes em dezembro para uma unidade no litoral, que recebeu em média 3.000 clientes por mês de jan a nov, tem nota média igual a 9 para os três aspectos avaliados e tempo de espera médio de 20'.

Arquivo: Satisfacao_Restaurante (.txt)

FIA BUSINESS SCHOOL lab.data

Referências Bibliográficas

REGRESSÃO LINEAR

- Kutner, M. H. et al. *Applied Linear Regression Models*. 4ª edição. McGraw-Hill Irwin, 2004.
- Rencher, A. C. e Schaalje, G. B. *Linear Models in Statistics*. 2ª edição, Wiley, 2008.
- Montgomery, D. C. *Introduction to Linear Regression Analysis*. 5ª edição. Wiley, 2012.
- James, G. *An Introduction to Statistical Learning - With applications in R*. 2ª edição. Springer, 2021.

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA