Regressão Linear Simples e Múltipla

Agora sim, vamos adentrar ao mundo de **Machine Learning** e partiremos justamente entendendo o processo de modelagem de dados.

Mas CALMA! O que exatamente significa modelagem?

Relaxe que a gente vai aprender agora. Vamos começar com um exemplo!

Imagina que você esteja procurando uma casa para comprar e, depois de analisar algumas, percebe que os preços oscilam de uma para outra. E até que faz sentido né? Isso porque variam os números de quartos, localização, se possui ou não piscina, e por aí vai!

E se quisermos saber o quanto cada uma dessas características altera o preço final da casa? É exatamente aqui que entra a parte de modelagem.

Modelar é escrever matematicamente a influência de variáveis na variável resposta. No exemplo acima, a variável preço é dependente das outras características da casa (variáveis independentes) e podemos “modelar” o comportamento do preço de acordo com cada uma delas.

Conclusão: utilizamos as variáveis independentes para “explicar” o preço de um determinado imóvel (nossa variável resposta).

Legal! Mas como podemos fazer isso?

Precisamos de ferramentas para descrever esse relacionamento entre as variáveis, certo? Para isso, as equações matemáticas serão nossas aliadas. Sim, exatamente aquelas que você viu nos tempos de escola e achou que nunca ia usar!

Lembra quando tínhamos um y que variava de acordo com um x? É esse o espírito da coisa! Isto é, quando queremos definir o comportamento de uma variável qualquer y (preço das casas

no nosso exemplo) através de uma variável x (tamanho do lote da casa, por exemplo), podemos escrever essa relação da seguinte forma:

$$y = a \cdot x + b$$

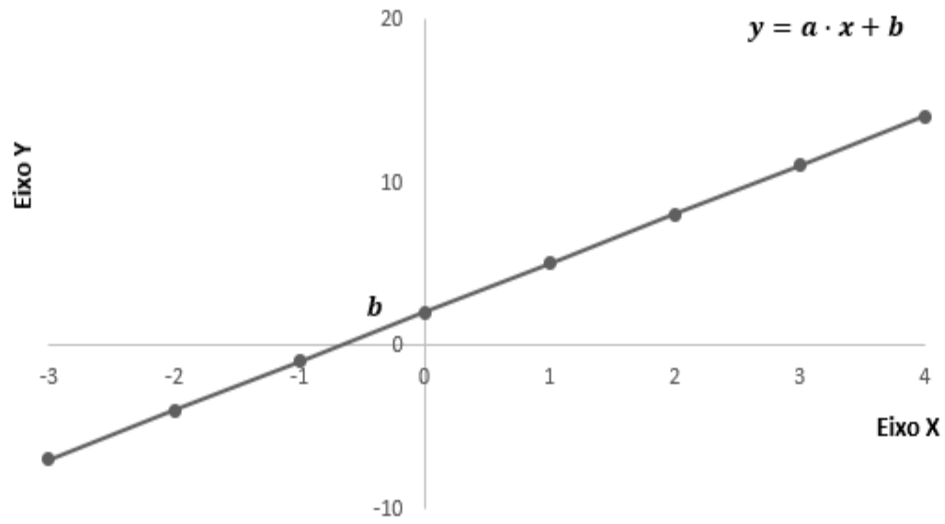
E aí, bateu saudades? Bora relembrar o que significa cada uma dessas letrinhas!

O y é a variável que depende de como x se altera, ou seja, y é dependente de x. Já x é a variável independente, que está sendo usada para explicar y.

O a é o coeficiente angular da reta, e ele define se a nossa reta será mais ou menos inclinada no gráfico.

Por fim, o b é o termo constante da reta, mais conhecido como intercepto. Esse valor nos mostra onde a reta vai encostar no eixo y.

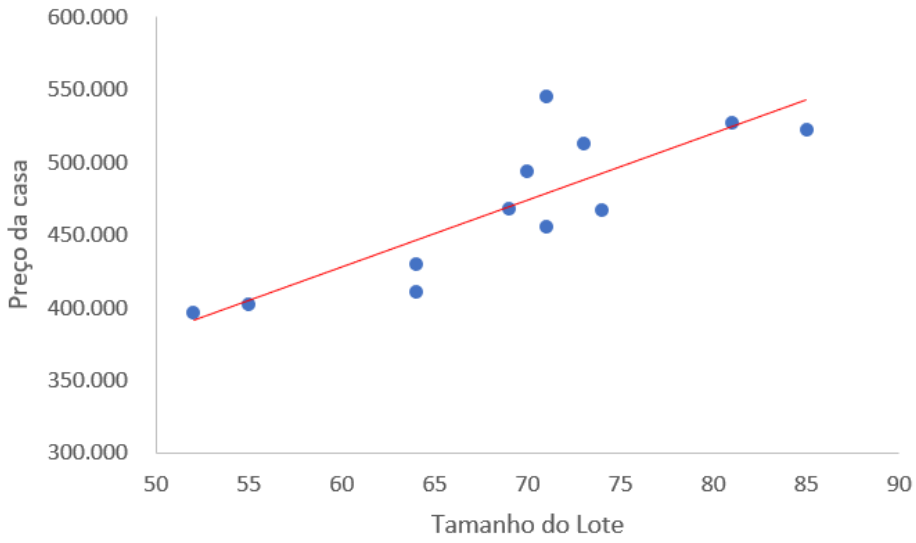
Graficamente, temos o seguinte:



Pronto, relembramos o suficiente para iniciarmos o estudo do nosso tão esperado primeiro modelo: **a regressão.**

E não, não revisamos a equação de reta a toa: o modelo de regressão descreve a relação entre as variáveis dependente e independente de forma linear, isto é, seguindo uma linha.

Voltando para o nosso exemplo, suponha que temos em mãos os preços de algumas casas e o tamanho do lote de cada uma delas. Analisando graficamente o comportamento de ambas as variáveis, podemos notar o seguinte:



Cada ponto azul no gráfico representa uma das casas, com seu respectivo tamanho de lote e preço final. Note que a linha em vermelho indica uma tendência linear dos pontos, isto é, eles tendem a se alinhar e a seguir essa reta.

Opa! Mas e esses pontos longe da reta?

Estamos analisando o comportamento dos pontos de uma forma geral, portanto é natural que alguns deles não estejam na reta e sequer próximos dela. Lembre-se que a modelagem sugere uma **tendência** que, no caso da regressão, é uma tendência linear.

Assim, é esperado que haja pontos que não sigam essa tendência. O objetivo não é atingir a perfeição (até porque na vida real isso é impossível), mas sim tentar chegar o mais próximo

possível do comportamento real dos dados, de forma que seja viável fazer previsões coerentes sobre eles.

Bom, conforme vimos no gráfico acima, foi possível entender melhor a relação entre as variáveis preço (variável dependente) e a variável tamanho do lote (variável independente). Quando relacionamos apenas duas variáveis (uma dependente e uma independente), estamos realizando uma análise do tipo **regressão linear simples**.

Da mesma forma, quando relacionamos a variável preço (variável dependente) com as variáveis tamanho do lote, número de quartos e localização (variáveis independentes), estamos fazendo uma análise do tipo **regressão linear múltipla** (ou seja, temos mais de uma variável que impacta a nossa variável dependente).

Bora recapitular esses conceitos?

Análise de Regressão: é o estudo da relação entre uma variável dependente (y) e outras variáveis independentes.

Regressão Linear Simples: é o estudo da relação entre uma variável dependente (y) e uma única variável independente (x).
Regressão Linear Múltipla: é o estudo da relação entre uma variável dependente (y) e duas ou mais variáveis independentes ($x_1, x_2, x_3, ..., x_n$).

Detalhe importantíssimo!

A regressão só pode ser utilizada para explicar variáveis quantitativas. Pensa só: estamos utilizando uma equação para explicar o comportamento de uma variável y, por isso ela precisa ser numericamente mensurável (possível de ser medida), isto é, possuir valores numéricos contínuos.

Um exemplo de variável contínua seria altura. Em um grupo de pessoas, podemos ter membros que medem 1,70m, 1,85m, 1,59m, e por aí vai. Veja que essa variável pode assumir

valores numéricos “quebrados” dentro de uma determinada faixa de valores possíveis, ao mesmo tempo que é possível medir a altura de cada indivíduo.

Agora, variáveis como número de filhos, por exemplo, não entram na categoria anterior. Apesar de serem numéricas, elas indicam uma contagem, e não uma medida. Nesse caso, são classificadas como variáveis quantitativas discretas, e não contínuas.
E lembre-se: o objetivo da regressão é ajustar uma equação que melhor explica o comportamento dos dados. Agora que entendemos os conceitos, vamos entrar um pouquinho mais a fundo na parte matemática da coisa?

Calma que você vai entender tudo!

Regressão Linear Simples

Vamos começar estudando mais profundamente a regressão linear simples. Nosso principal objetivo será o de supor uma equação linear que explique como a uma variável independente interfere no valor de uma variável dependente.

Lembra da equação de primeiro grau que revisamos anteriormente? Pois bem, ela também é utilizada na regressão, porém com “letrinhas” diferentes. Saca só:

$$y_n = \beta_0 + \beta_1 \cdot x_n + e_n$$

O valor y na esquerda da equação é o valor real dessa variável, a que observamos no mundo real. Já do lado direito, $\beta_0 + \beta_1 \cdot x_n$ corresponde à aproximação linear da variável y que estamos fazendo através da regressão linear, isto é, o valor de y estimado. E por se tratar de uma aproximação, incluímos o erro aleatório ou resíduo e_n.

Esse erro representa todas as influências no comportamento da variável y que não podem ser explicadas linearmente pela variável x. Isto é, que não conseguimos explicar através do nosso modelo linear.

Como sabemos que nosso modelo não é perfeito, precisamos incluir os erros também, não é mesmo?

Perceba que o intercepto da equação que vimos anteriormente era b, e agora ele é representado através do símbolo β_0. Já o termo que multiplicava x na primeira equação era a, e agora ele passou a ser representado por β_1.

Veja que y_n é o valor observado da variável y, da mesma forma que x_i é o valor observado da variável x. Logo, nossos dados são da forma de pares ordenados (x_n, y_n), ou seja, para cada observação de x existe um valor observado de y (por exemplo, para cada tamanho de lote observado nos dados, há o preço correspondente da casa localizada neste lote).

Importante!

Esses "n" presentes nas variáveis indicam os n possíveis valores que a variável dependente y pode assumir, e o mesmo vale para x (por exemplo $x_1 = 1,\ x_2 = 2,\ ....,\ x_{10} = 300,\ n = 1, 2, 3, ..., 10.$). Tranquilo, né?

Beleza, mas o que são esses "betas" (β) na equação?

Esses betas representam os coeficientes de regressão, isto é, são os parâmetros desconhecidos do modelo e que também precisam ser estimados. O β_0 representa o valor esperado de y quando a variável independente é nula (x = 0). Já o β_1 representa o quanto varia em y para cada mudança de uma unidade em x.

Detalhe importante:

Nem sempre o intercepto (ou no caso da equação, β_0) fará sentido no contexto real do problema estudado. Afinal, é extremamente raro que todas as variáveis independentes do modelo se anulem ao mesmo tempo, ou até mesmo que algumas assumam valores nulos.

Por exemplo, ao analisar preços de imóveis, não faz sentido ter um lote com zero metros quadrados, não é mesmo?

O intercepto tem uma finalidade mais matemática do que de fato prática (na vida real). De forma resumida, ele garante o ajuste adequado da linha de regressão com objetivo de minimizar erros de previsão do modelo.

Para reforçar, preste atenção: **nós nunca obteremos uma equação exata que modela um determinado conjunto de dados.** Isso é inviável no mundo real. Porém, conseguimos chegar próximo de uma equação que explique bem esse conjunto através da regressão.

Observação Importante!

É extremamente raro que tenhamos os dados de toda a população de estudo do problema, o que inclusive justifica a aplicação da inferência estatística durante a análise de dados. Nesse caso, retiramos uma parte do todo (amostra) e, a partir dela, tiramos conclusões sobre a população inteira.

Por isso, é mais comum nos depararmos com a função de regressão amostral, que é justamente a mesma equação vista acima, porém com “chapeuzinhos” nas variáveis. Eles indicam apenas que essas variáveis com “chapéu” foram estimadas a partir de uma amostra .

Saca só como fica a equação:

$$y_n = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_n + \hat{e}_n$$

De boa né?

O lado esquerdo, conforme já vimos, é o valor da variável dependente observado na realidade. Já o lado direito contém a aproximação dela via regressão linear, junto com o erro que acompanha essa aproximação.

Portanto, nosso y estimado pode ser escrito da seguinte maneira:

$$\widehat{y}_n = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_n$$

Beleza, mas como verificar se nossa aproximação é boa afinal?

A melhor forma de fazer isso é verificar o quanto a nossa aproximação chega perto do valor verdadeiro. Para fazer isso, vamos simplesmente subtrair o valor real da variável y do valor aproximado que calculamos através da regressão.
Lembra que a equação da regressão incluía também os possíveis erros que poderíamos obter? Então, é através deles que iremos comprovar se o modelo é de fato eficaz. Saca só:

$$y_n = (\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_n) + \widehat{e}_n$$

Olha lá! A equação em parênteses é exatamente a equação do nosso y estimado. Assim, temos o seguinte:

$$y_n = \widehat{y}_n + \widehat{e}_n$$

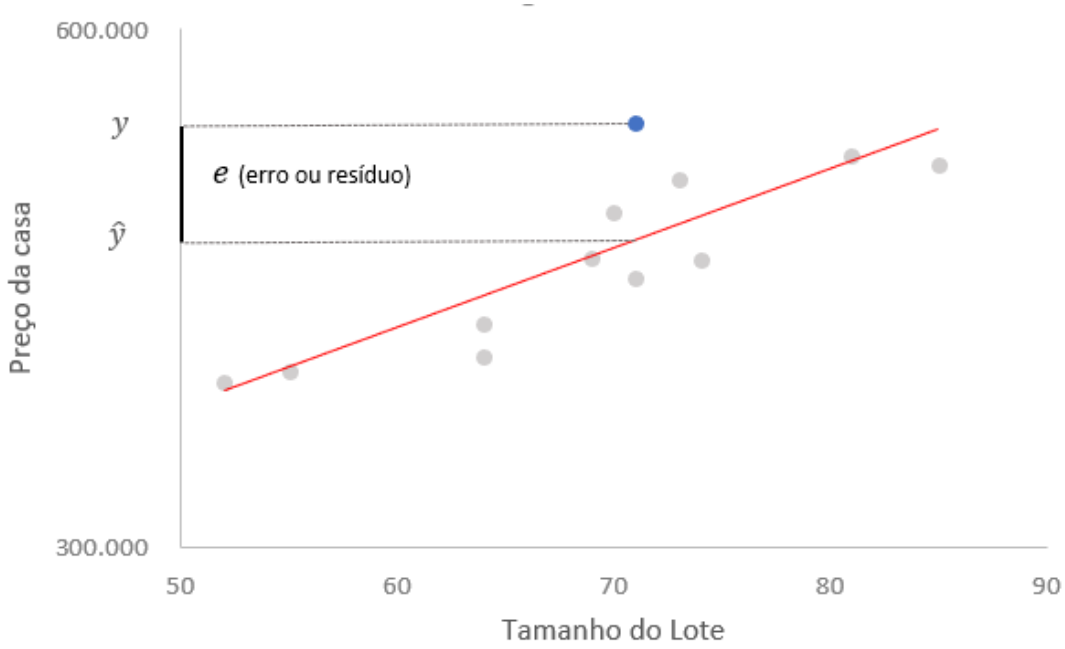
ou

$$y_n - \widehat{y}_n = \widehat{e}_n$$

Veja só! Fazendo a diferença entre o valor real observado e o valor estimado (y e $\widehat{y}$ respectivamente), obtemos justamente aquele erro que já discutimos!

Mas o que seriam exatamente esses erros, hein?

Vamos ver isso no próprio gráfico, saca só:



Vamos selecionar apenas um dos pontos do gráfico (de cor azul) para exemplificar!

É possível observar que o ponto azul representa o valor real de y, que pelo gráfico dá mais ou menos 70 (tamanho do lote). Enquanto isso, para esse mesmo valor de x, ou seja, para o mesmo tamanho de lote, a reta estimou um valor ($\hat{y}$), um pouco abaixo do valor real y.

No eixo de "preço da casa", dá pra ver que essa distância entre y e $\hat{y}$ representa o erro ou resíduo que discutimos anteriormente.

E nossa missão, qual é?

É de justamente **minimizar esses erros**, para que nossas previsões fiquem bem próximas dos valores reais que y assume. Assim, somando todos esses erros, conseguimos obter o total de informação que nosso modelo linear falhou em explicar.

Epa, mas tem um probleminha aí né?

Veja que, no gráfico, que alguns pontos de dados estão abaixo da reta de regressão, levando a uma diferença negativa entre y e $\hat{y}$. Erro negativo não rola né? Até porque, na hora de somar

todos os erros obtidos, eles poderiam se anular ou trazer uma soma incoerente com a realidade.

Como resolver esse problema?

Para isso, temos uma saída melhor: elevar cada um dos erros ao quadrado, e só depois somar todos eles. Assim, não corremos o risco de obter uma soma incorreta. Genial, né? É exatamente isso que faz o famoso **Método dos Mínimos Quadrados (MMQ)**.

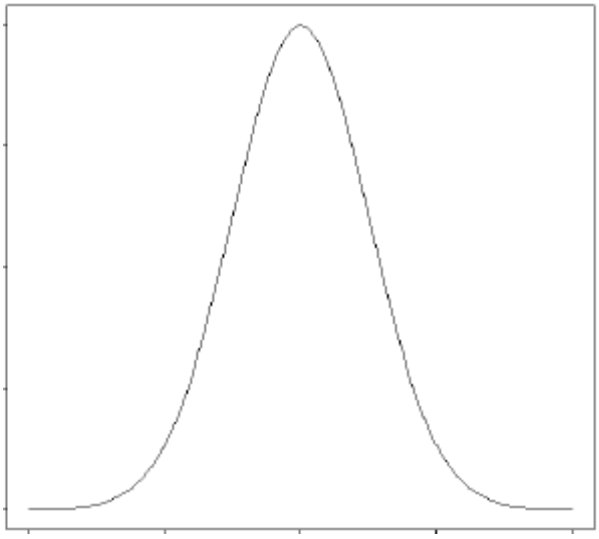
Somando o quadrado de todos os erros obtidos pelo modelo, nosso objetivo será, agora, o de minimizar o valor dessa soma. Afinal, quanto menor o valor dela, significa que menores são os erros obtidos para cada um dos valores previstos pelo modelo de regressão.

Perceba que, por se tratar de uma soma dos desvios (erros) ao quadrado, esse método acaba sendo bastante sensível quanto à presença de **outliers**, que promovem erros maiores do que o esperado para o modelo e, consequentemente, elevam o valor da soma. Esse efeito é diretamente refletido na performance do modelo, que tende a piorar com a presença de outliers.

Ah! É através do MMQ que estimamos os valores dos betas da nossa equação linear. Afinal, eles são decisivos para nos ajudar a minimizar essa bendita soma através do melhor ajuste da linha de regressão.

Comportamento dos erros

A análise de regressão estabelece que os erros (resíduos) devem ser distribuídos conforme a curva normal, com média zero e variância comum. Essa distribuição você já estudou, mas não custa relembrar o formato dela. Saca só:



(Fonte: Wikipedia)

O fato de se desejar uma média zero para os resíduos significa apenas que o desejado é que a média dos erros obtidos com o modelo seja próxima de zero. Afinal, sabemos que nosso modelo vai errar, mas o ideal é que esses erros sejam pequenos, não é mesmo?

A condição de variância comum estabelece que os erros devem se distanciar de uma maneira "parecida" da média zero. Isso significa que os resíduos devem possuir uma variabilidade bastante próxima, e não seguir tendências de comportamento.

Essa condição é formalmente conhecida como condição de **homocedasticidade** (o oposto dela é conhecido como heterocedasticidade). O nome é assustador, mas a ideia é exatamente o que discutimos: obedecer à condição de variância comum.

A dispersão dos erros também ajuda a identificar se o modelo linear é de fato adequado para os dados ou não. Quanto menos homogênea for essa dispersão, investigue se mais variáveis podem ser adicionadas, ou se é melhor utilizar modelos polinomiais, por exemplo.

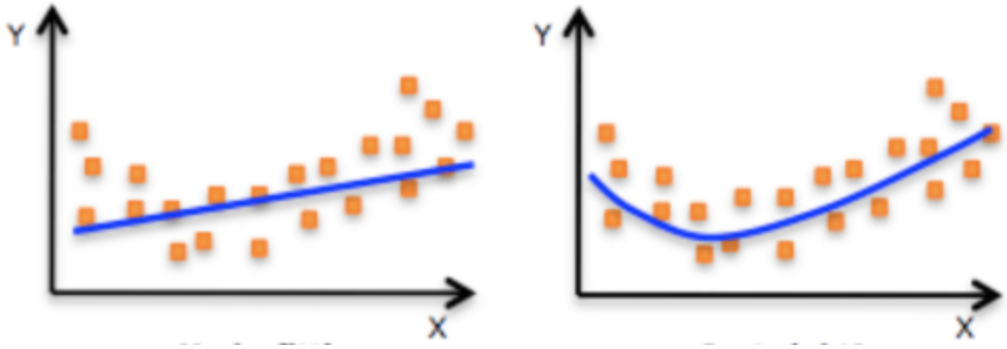
O motivo de se desejar uma dispersão homogênea para os erros é o seguinte: o modelo é tão eficiente para explicar o comportamento da variável resposta que apenas fenômenos aleatórios (existentes em qualquer fenômeno da vida real) são responsáveis pelos erros obtidos. Isto é, o modelo só erra porque é impossível de prever aleatoriedade.

No entanto, a falta de homogeneidade indica que as variáveis independentes do modelo não conseguem explicar completamente o comportamento da variável resposta.

Essa parte não explicada acaba se refletindo nos resíduos e muitas vezes acontece pela ausência de alguma variável explicativa, ou até mesmo a ausência de um termo de ordem superior em alguma variável que explique melhor a curvatura dos dados.

É comum termos problemas que incluem efeitos não lineares e, para resolver isso, podemos incluir a mesma variável no modelo, tanto na forma linear (elevada a 1) quanto na forma não linear (elevada ao quadrado, ao cubo, etc). Isso pode ser feito principalmente quando a real curvatura dos dados está distante da linha de regressão do modelo.

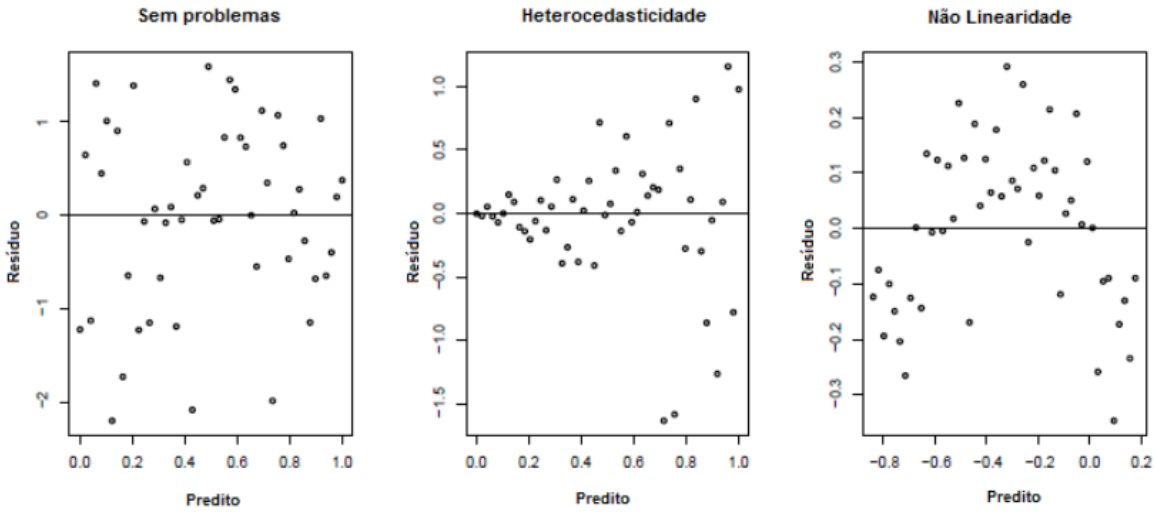
Veja um exemplo gráfico dessa situação:



(Fonte: https://www.datarobot.com/wiki/underfitting/)

A reta do gráfico esquerdo não explica bem o comportamento real dos dados. Esse efeito não linear pode ser representado melhor pela curva da direita, que certamente inclui alguma variável elevada a ordens superiores.

Para ficar mais claro, veja os gráficos abaixo com possíveis dispersões dos resíduos:

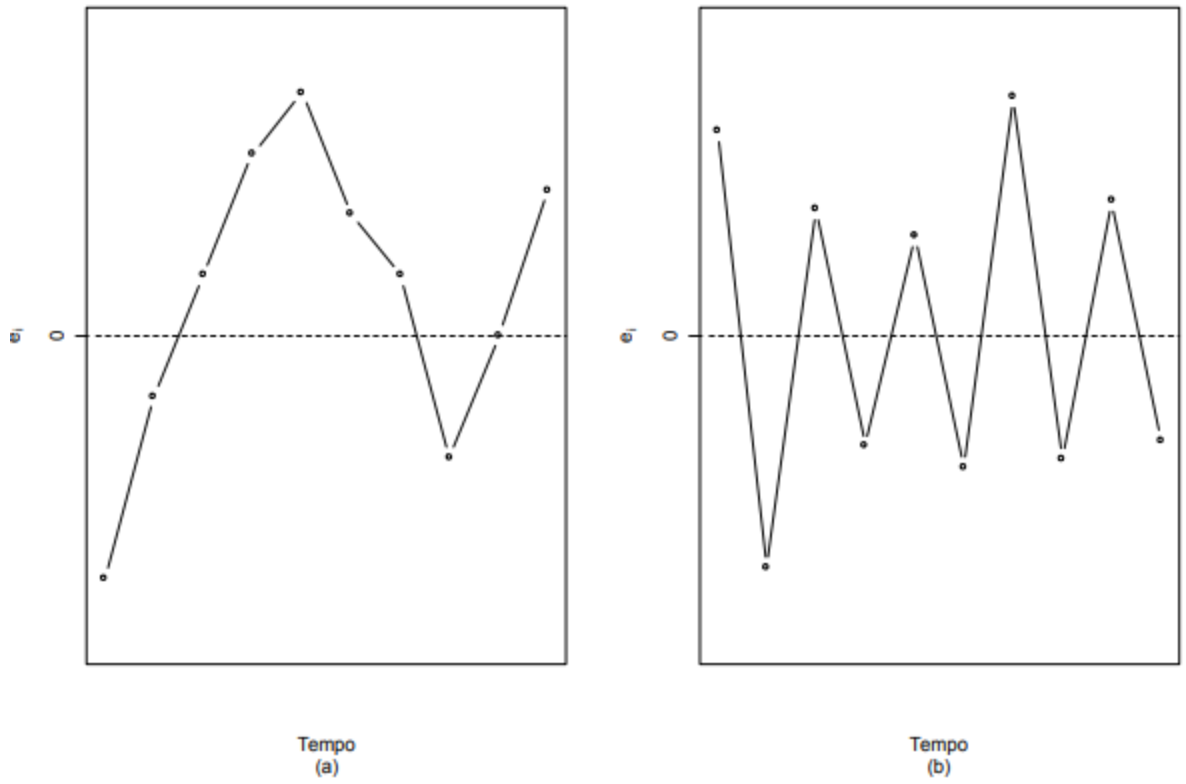


(Fonte: UFPR)

Outro cuidado importante é o de verificar se os erros não estão correlacionados entre si. Isto é, eles não podem seguir nenhuma tendência específica que indique qualquer tipo de associação do erro anterior com o próximo.

Esse fenômeno é bastante comum quando lidamos com dados ordenados por tempo, por exemplo. Uma vez que cada resposta depende da anterior, os erros acabam refletindo esse comportamento também. Nesse caso, é melhor avaliar outros tipos de modelagem, como análise de séries temporais (essenciais para lidar com variações no tempo).

Veja só alguns exemplos disso:



(Fonte: UFPR)

Em resumo, se os resíduos do seu modelo não estiverem distribuídos de uma maneira aleatória e sem tendências…..**SUSPEITE**!

Bora revisar o que acabamos de aprender?

Método dos Mínimos Quadrados: é uma técnica de otimização que procura encontrar o melhor ajuste para um conjunto de dados, de forma a minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados. É através desse método que se calculam os betas que minimizam os erros e, consequentemente, a soma dos erros quadráticos.

Avaliação da performance do modelo

Existem vários passos que nos guiam para a avaliação do modelo de regressão (intervalos de confiança para os parâmetros estimados, testes de hipóteses para descarte ou não de parâmetros, tabela ANOVA, dentre outros). Entretanto, vamos focar no **coeficiente de determinação** (r^2)**.**

Esse coeficiente indica a proporção da variação dos dados que é explicada pelo modelo de regressão. Isto é, o quanto de fato nosso modelo consegue explicar o comportamento dos dados em questão.

O coeficiente de determinação pode obter os seguintes resultados:

$r^2 \approx 0$ **: Modelo praticamente não explica o comportamento dos dados**

$0 \leq r^2 \leq 1$ **: Modelo pode ou não ser adequado**

$r^2 \approx 1$ **: Modelo explica (em teoria) muito bem o comportamento dos dados**

CUIDADO!

Valores muito próximos de 1 para o coeficiente de variação podem indicar que o modelo de regressão linear está sofrendo overfitting. Isto é, o modelo não consegue generalizar bem o comportamento dos dados e, ao ser validado, pode ter um desempenho bastante ruim. Veremos isso em detalhes no próximo material, não se preocupe!

Não existe uma definição de valor ideal desse coeficiente, mas é preciso observar se o valor obtido é coerente e, posteriormente, validar o modelo com novos dados. Lembre-se: não existe modelo perfeito!

Regressão Linear Múltipla

Já tá craque em regressão simples? Então partiu entender mais sobre a regressão múltipla! Na real, a principal mudança na regressão linear múltipla é a inclusão de mais variáveis independentes na explicação do comportamento da variável dependente em estudo.

No nosso exemplo, poderíamos incluir, além da variável “tamanho do lote”, a variável “número de quartos” para auxiliar na modelagem do comportamento dos preços das casas.

Assim, nosso modelo poderia ser representado da seguinte forma:

$$y_n = \beta_0 + \beta_1 \cdot x_{1_n} + \beta_2 \cdot x_{2_n} + e_n$$

Nesse caso, a variável “número de quartos” possui o coeficiente β_2 e é representada por x_2 na equação acima. Seguindo essa ideia, a equação geral para um modelo de regressão múltipla seria:

$$y_n = \beta_0 + \beta_1 \cdot x_{1_n} + ... + \beta_k \cdot x_{k_n} + e_n$$

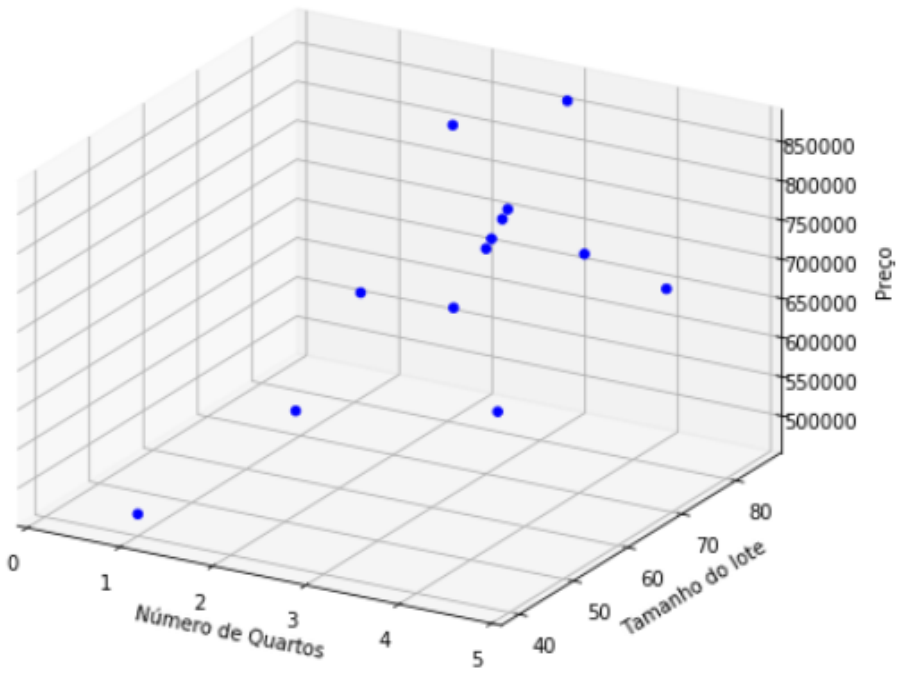
Perceba que agora temos duas variáveis independentes que explicam y, e que os betas aumentam conforme adicionamos novas variáveis. Isso é esperado, pois queremos verificar exatamente o quanto cada uma das variáveis independentes interfere, de forma isolada, na variável dependente.
É importante destacar que cada beta é calculado mantendo-se as demais variáveis incluídas constantes. Assim, obtemos a variação esperada em y para cada variação unitária da variável x em questão.

Opa! Mas como fica a cara do gráfico nesse caso?

Como aumentamos o número de variáveis independentes no nosso modelo, aumentamos também a complexidade dele.

Como assim?

Aumentamos a dimensão do nosso modelo, pois agora y varia de acordo com 2 variáveis (tamanho do lote e número de quartos no nosso exemplo). Saca só:



Portanto, ao adicionar cada vez mais variáveis, vamos alcançando novas dimensões, e é exatamente por isso que você estudou um pouco de álgebra linear! Afinal, vamos precisar de métodos mais sofisticados para fazer os cálculos, principalmente envolvendo matrizes.

Mais uma vez, não se preocupe com essa parte mais complexa do conteúdo, mas é importante entender o motivo dos cálculos. O que diferencia a regressão múltipla é apenas **a quantidade de variáveis independentes envolvidas na explicação da variável dependente**.

Continuaremos utilizando o Método dos Mínimos Quadrados e o coeficiente de variação (ambos explicitados anteriormente) para verificar a adequação do modelo linear perante uma base de dados qualquer.

Bizu da regressão:

Bora dar aquela revisada rápida no que aprendemos hoje:

Regressão Linear: é um método que tenta explicar uma determinada variável (variável dependente) através de outras variáveis (variáveis independentes).

- **Regressão Linear Simples:** apenas uma variável independente (x) é considerada para modelar o comportamento da variável dependente (y);
- **Regressão Linear Múltipla:** duas ou mais variáveis independentes ($x_1, x_2, ..., x_n$) são utilizadas para modelar o comportamento da variável dependente (y).

Método dos mínimos quadrados: é um método que consiste em encontrar os melhores valores para os parâmetros (os betas) da equação do modelo de regressão. O objetivo é minimizar a soma dos quadrados dos erros obtidos e obter a melhor aproximação linear possível para os dados em questão.

Coeficiente de determinação (r^2 ou r-quadrado): indica o quanto o modelo de fato conseguiu "explicar" os dados através da equação linear. Isto é, uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. Ele varia conforme o seguinte:

$r^2 \approx 0$ **: modelo não explica nada da variabilidade dos dados**
$0 \leq r^2 \leq 1$ **: modelo pode ou não ser eficiente**
$r^2 \approx 1$ **: modelo explica quase que completamente a variabilidade dos dados**

Lembrete!

Lembre-se sempre de explorar bem a base de dados e, principalmente, de entender o problema que se deseja resolver e a natureza dos dados em questão.

Um r-quadrado muito alto nem sempre indica um modelo muito bom, pois pode ocorrer perda de generalidade do modelo (o temido overfitting).

Já um r-quadrado baixo também nem sempre indica que o modelo é ruim. Em algumas áreas de estudo o valor esperado desse coeficiente é baixo, pois alguns comportamentos são difíceis de prever.

Dica:

Tente sempre **plotar e analisar o gráfico dos resíduos obtidos.** A regressão impõe que os resíduos possuam distribuição normal com média zero, então verifique sempre o comportamento deles!

Além disso, faça uma análise descritiva bastante profunda dos dados. Ajuda a ter insights sobre o comportamento deles e sobre o modelo que deve ser implantado!

E aí, aprendeu?

Quer saber mais?
Dá uma olhada nos links abaixo:

Curso online - Duke University:
Linear Regression and Modeling

Série de vídeos do StatQuest:
Linear regression and Linear Models