Aprendizado de Máquina

Parte II

Regressão Linear

Regressão linear simples: $Y = a + bX + u$

Regressão linear multivariada:

$Y = b + a_1 X_1 + a_2 X_2 + \ldots + a_d X_d + u$

$onde,$

Y : Variável dependente a ser predita, classificada ou explicada.

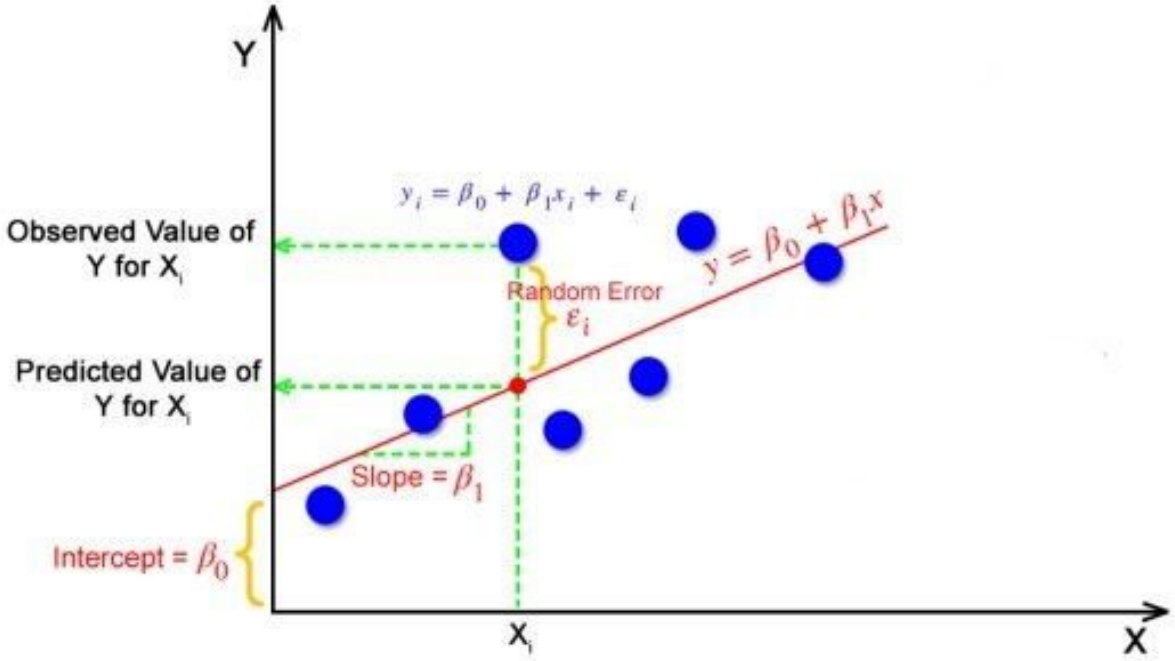
X ou X_i : Variável independente usada na determinação de Y.

b : Coeficiente linear

a ou a_i : Coeficiente angular da variável independente

u : regressão residual ou erro

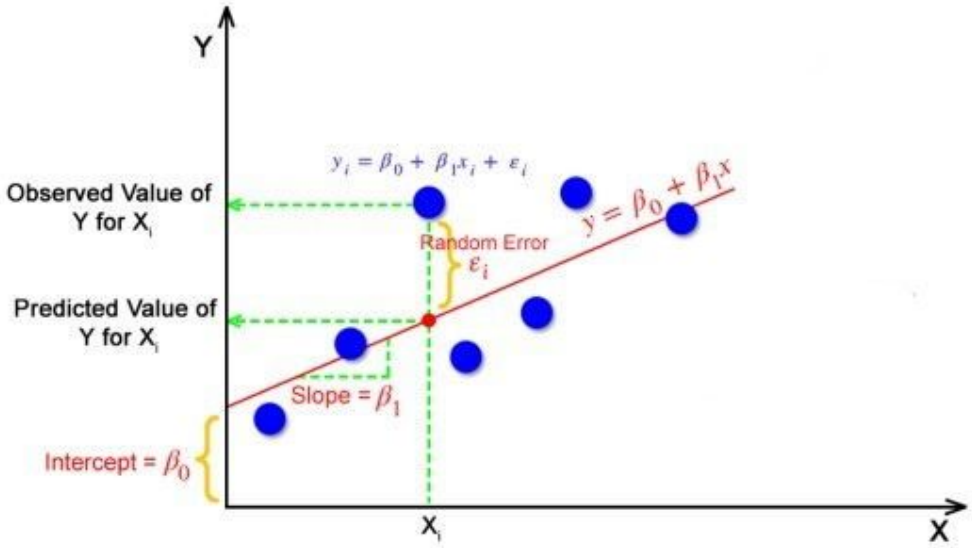
Regressão Linear



Regressão Linear

Algumas hipóteses consideradas até agora:

- O conjunto de dados X apresenta atributos (variáveis) independentes.
- Há uma dependência do atributo-alvo Y em relação ao conjunto de dados X.
- Essa relação entre X e Y é linear.



Regressão Linear

- Um modelo de regressão linear simples vai estimar o coeficiente angular e o coeficiente linear da reta que melhor representa (best fit) a relação entre os atributos (variáveis independentes) e o atributo-alvo (variável dependente).
- O coeficiente angular indica a mudança na variável dependente para cada unidade de mudança na variável independente.
- O coeficiente linear indica o valor predito da variável dependente quando a variável independente é zero.

Regressão Linear

- A chamada best fit line será a reta que minimiza a soma do erro residual quadrático, ou Residual Sum of Squares (RSS).
- Utiliza-se o Mean Squared Error (MSE) ou erro quadrático médio como métrica a ser minimizada.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

- Temos um problema de otimização

$$min\frac{1}{N}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Regressão Linear

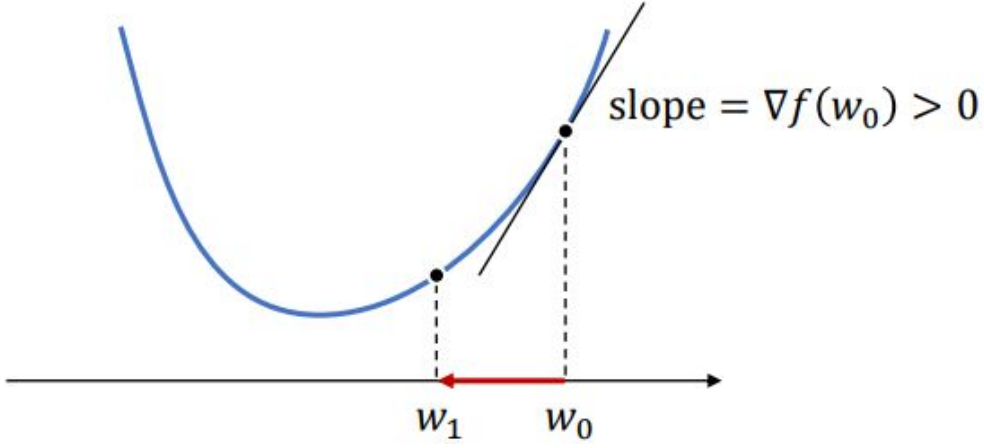
- Podemos aplicar o algoritmo Gradiente Descendente

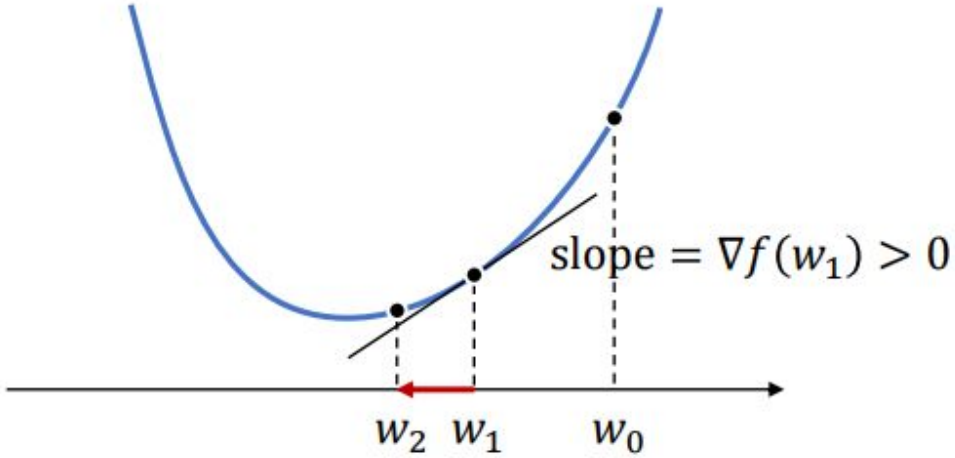
Passo 1: Inicie os pesos aleatoriamente

Passo 2: Calcule os gradientes dos parâmetros da função de custo f(y,x) - $\partial f(W)/\partial W = \nabla f(w)$.

Passo 3: Atualize os pesos W = W - ηf

Passo 4: Repita até o custo f(w) parar de reduzir, ou algum outro critério de parada seja satisfeito.





Regressão Linear

- Podemos aplicar o algoritmo Gradiente Descendente

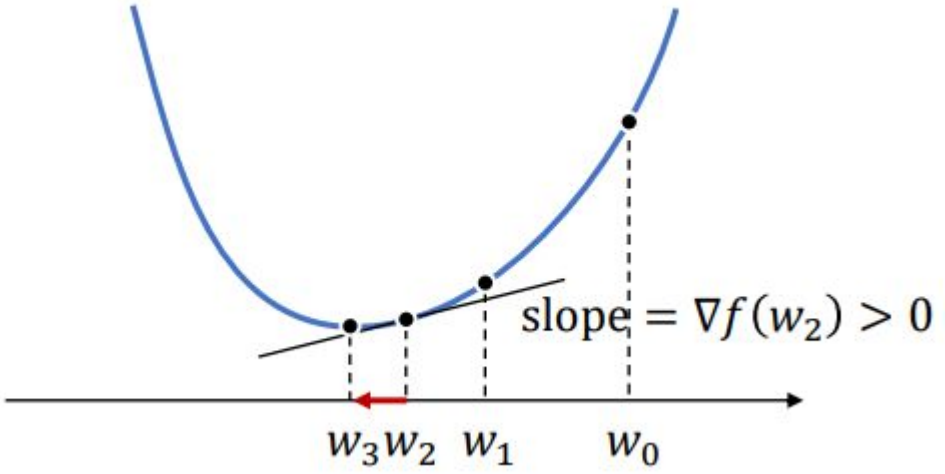
Passo 1: Inicie os pesos aleatoriamente

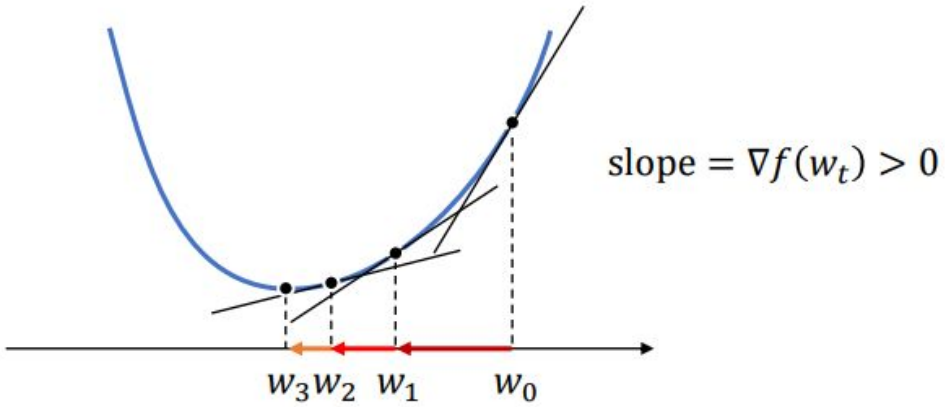
Passo 2: Calcule os gradientes dos parâmetros da função de custo f(y,x) - $\partial f(W)/\partial W = \nabla f(w)$.

Passo 3: Atualize os pesos

$W = W - \eta \nabla f(w)$

Passo 4: Repita até o custo f(w) parar de reduzir, ou algum outro critério de parada seja satisfeito.





Regressão Linear

$$f = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

$$\frac{\partial f}{\partial a} = \frac{2}{N} \sum_{i=1}^{n} (a + bx_i - y_i)$$

$$\frac{\partial f}{\partial b} = \frac{2}{N} \sum_{i=1}^{n} (a + bx_i - y_i) . x_i$$

$$a = a - \eta(\frac{2}{N} \sum_{i=1}^{n} (a + bx_i - y_i))$$

$$b = b - \eta(\frac{2}{N} \sum_{i=1}^{n} (a + bx_i - y_i) . x_i)$$

Regressão Linear

- As métricas mais utilizadas para avaliar a adequação de um modelo de regressão linear são:
  - R-Squared (R2)
  - Root Mean Squared Error (RSME)
  - Residual Standard Error (RSE)

Regressão Linear

- R-Squared (R^2): avalia o qual próximo os atributos estão do modelo de regressão

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS : Residual sum of Squares

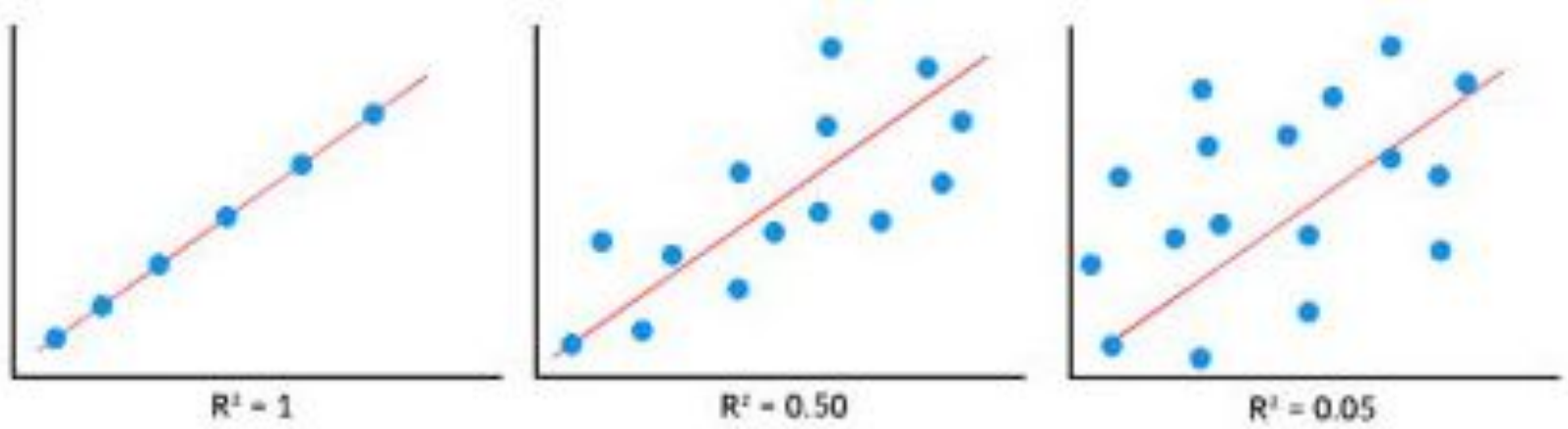
$$RSS = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

TSS : Total Sum of Squares

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

Regressão Linear

- $R^2 \in [0,1]$, onde quanto mais próximo de 1, melhor o modelo está ajustado aos atributos.



Regressão Linear

- Root Mean Squared Error (RMSE): raiz quadrada da variância dos resíduos
- Determina quão próximo os dados estão dos valores preditos.

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{dado} - y_i^{previsto})^2}$$

- RMSE sofre influência das unidades das variáveis (não é uma métrica normalizada), ou seja, isso significa que pode variar dependendo da unidade das variáveis.

Regressão Linear

- Residual Standard Error (RSE): deixa a métrica menos enviesada ao dividir pelos graus de liberdade ao invés do total de dados.

$$RMSE = \sqrt{\frac{RSS}{df}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i^{dado} - y_i^{previsto})^2}$$

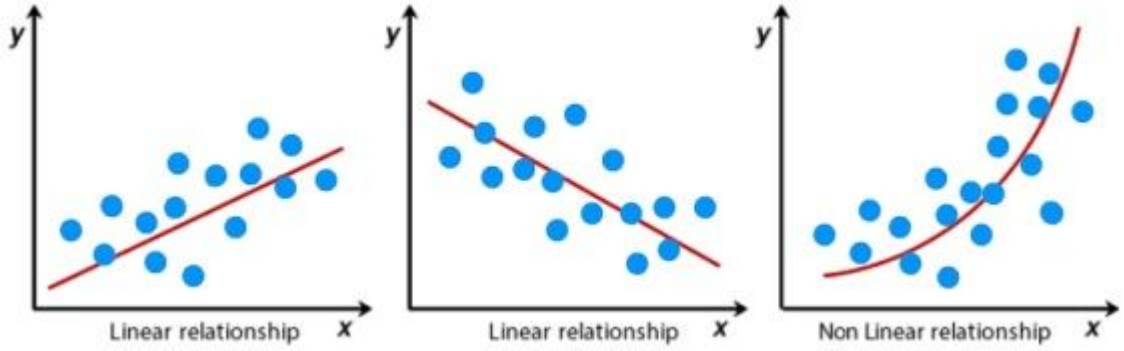
- R-squared se torna melhor por trabalhar com valores normalizados.

Regressão Linear

- Algumas hipóteses devem ser satisfeitas sobre os dados para uma análise satisfatória usando regressão linear.
- Ausência de multicolinearidade: sem correlação alta entre variáveis independentes.

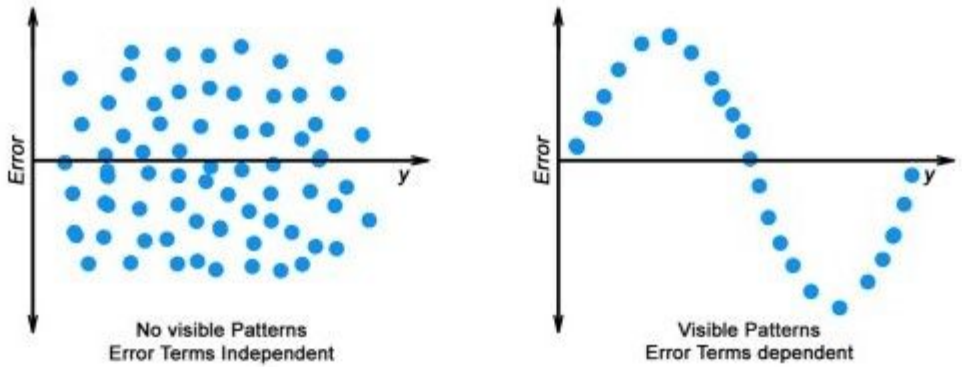
Regressão Linear

- Linearidade: relação linear entre x_i (dados independentes) e y_i (dados dependentes).
- Mudanças em x_i levam a mudanças em y_i de forma linear



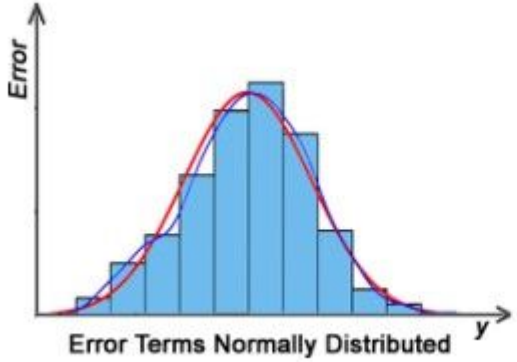
Regressão Linear

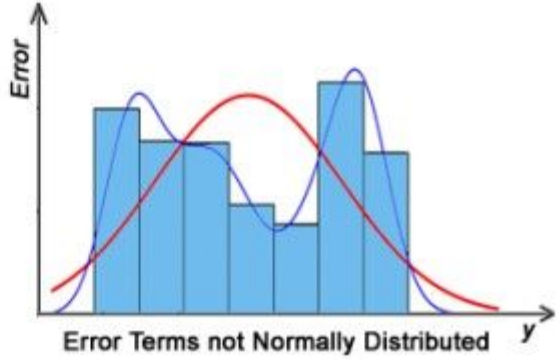
- Independência: os atributos-alvo são independentes, ou seja, o valor de uma variável dependente não está correlacionado ao valor de outra variável dependente.
- Visualmente não se consegue identificar um padrão nos valores de y_i



Regressão Linear

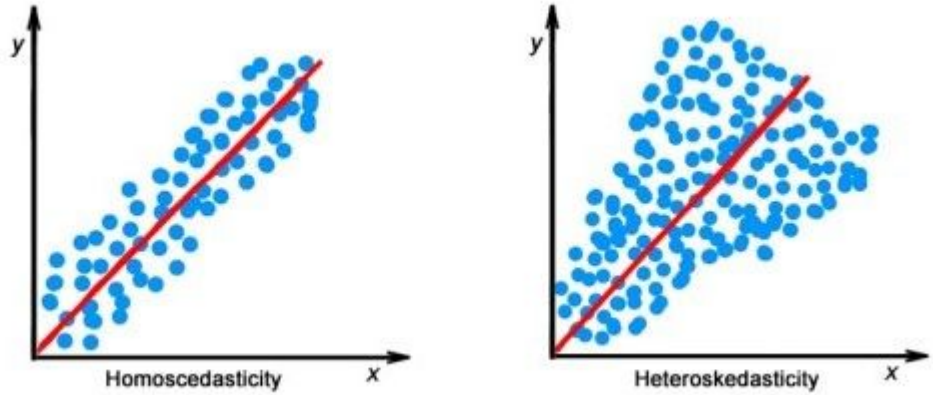
- Resíduos seguem uma distribuição Normal com média zero ou próxima a zero.
- Verifica-se desta forma se a regressão atual é a que melhor se adequa ou não aos dados.
- Resíduos que não seguem a distribuição normal indicam a existência de dados que precisam ser avaliados para melhor ajuste do modelo.





Regressão Linear

- Homocedasticidade: os resíduos seguem uma variância constante.
- A não constância da variância, chamada Heterocedasticidade, indica a existência de outliers ou valores extremos.



Regressão Linear

- Teste de hipótese para o modelo:

  A reta obtida se adequa de forma significativa aos dados?”

  ou

  O coeficiente b explica a variação nos dados?

  - H_0: b = 0
  - H_1: b ≠ 0

Regressão Linear

- Podemos aplicar os testes abaixo para verificar o quanto o modelo está adequado aos dados:
  - t-test: compara médias entre dois grupos de valores, verificando se as médias dos dois grupos são significativamente diferente.
  - Determina o p-value que mede a probabilidade de se obter o valor observado assumindo que a hipótese nula é verdadeira.
  - Logo, determina se o coeficiente angular é significativo ou não.

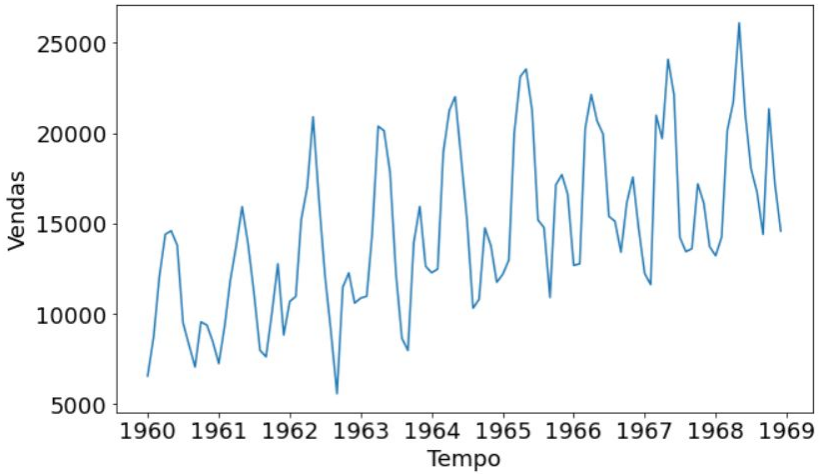
Regressão Linear

- F-test: compara variâncias ou médias para três ou mais grupos ou condições.
    - Determina se há uma diferença significativa na variância ou média entre vários grupos.
    - Logo, determina se o ajuste geral do modelo é significativo ou não.
    - Valores alto de F indicam um modelo mais ajustado.

ARIMA

ARIMA

- Uma **série temporal** descreve uma sequência de dados em diferentes intervalos de tempo (diário, mensal, anual).
- A **previsão em série temporal** aplica modelagem estatística para predizer valores futuros de uma série temporal baseado em dados anteriores na linha temporal.



ARIMA

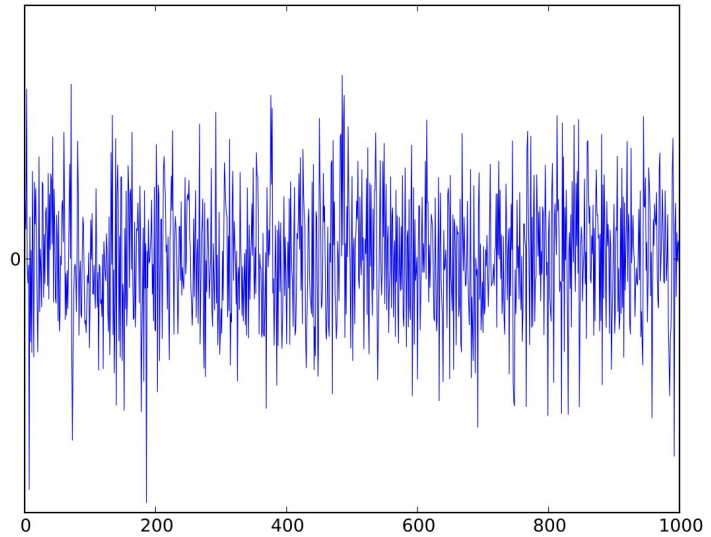
- Temos uma previsão em série temporal univariada quando usamos apenas os dados anteriores da série temporal para prever os valores futuros.
- Se dados externos (variáveis exógenas) são também utilizadas na predição, temos uma previsão multivariada em série temporal.

ARIMA

- Autoregressive Integrated Moving Average Model (ARIMA): explica uma dada série temporal a partir dos seus valores passados.
- Para isso, considera dados fora de fase e os respectivos erros previstos nesses casos.
- O Modelo também pode ser usado para previsão de valores futuros.

ARIMA

- Requisitos para aplicar ARIMA: séries não sazonais seguindo algum tipo de padrão e séries que não sejam ruído branco aleatório.
- Uma **série ruído branco** apresenta uma sequência de valores aleatórios que não podem ser preditos.
- Nesse caso, as variáveis são independentes e identicamente distribuídas com média zero, ou seja, apresentam mesma variância e zero correlação com outros valores da série.



ARIMA

- O modelo ARIMA é especificado por três parâmetros de ordem (p,d,q) usados em três componentes:
  - AR(p) - Autoregression
  - I(d) - Integration
  - MA(q) - Moving Average

ARIMA

- AR(p) Autoregression: utiliza p valores passados (*lags*) para estabelecer a equação de regressão da série temporal.
- Logo, baseia-se na relação de dependência entre o valor atual e os valores dos períodos anteriores.
- O valor p fornece a ordem da regressão, ou seja, o número de valores anteriores usados como preditores.

$$Y_t = b + a_1 Y_{t-1} + a_2 Y_{t-} + \ldots + a_p Y_{t-p} + \epsilon_t$$

ϵ_t : Erro do modelo AR ou ruído branco

$b, a_1, a_2, \ldots, a_p$: Parâmetros estimados pelo modelo

ARIMA

- I(d) Integration: subtrair um valor do valor anterior buscando tornar a série temporal estacionária.
- Uma série temporal pode ser dita um processo fracamente estacionário ou estacionário de segunda ordem quando:
  - A covariância depende unicamente da distância temporal entre seus valores
  - Apresentar média e variância constantes,
- Nesse caso, o processo mantém suas propriedades (padrão) ao longo do tempo.

Arima

- Matematicamente temos que uma série temporal é estacionária quando:

$$1. E(Y_t) = \mu$$
$$2. Var(Y_t) = \sigma^2$$
$$3. Cov(Y_t, Y_{t+k}) = f(k)$$

ARIMA

- Diferenciar a série consiste em subtrair o valor atual dos seus valores anteriores d vezes.
- Trata-se de uma transformação via diferenciação estatística realizada nos dados de uma série temporal para torná-la estacionária.
- Tal transformação se justifica já que as propriedades de uma série temporal estacionária não dependem do tempo em que a série está vinculada.

$$
\begin{aligned}
Y'_t &= Y_t - Y_{t-1} && \text{Primeira diferença} \\
Y''_t &= Y'_t - Y'_{t-1} && \text{Segunda diferença} \\
&= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\
&= Y_t - 2Y_{t-1} - Y_{t-2} \\
Y'_t &= Y_t - Y_{t-s} && \text{Diferença sazonal}
\end{aligned}
$$

ARIMA

- Os parâmetros de uma série serão afetados por um excesso de diferenciação da série (over-difference).
- Uma série com over-difference pode ser estacionária a custa de tais parâmetros.
- O correto é diferenciar o mínimo de vezes possível para alcançar uma série aproximadamente estacionária.
- Isso pode ser estabelecido pelos valores obtidos na média ou visualmente.

ARIMA

- Se as autocorrelações são positivas para 10 ou mais valores, a diferenciação poderia ser aplicada.
- Se as autocorrelações são muito negativas, a série pode ter over-difference.
- Na dúvida entre duas possíveis ordens de diferenciação, utilize aquela que alcança o **menor desvio padrão**.

ARIMA

- MA(q) Moving Average: calcula a média móvel a partir dos valores prévios dos **erros residuais**, onde q indica o número de termos incluído no cálculo.

$$Y_t = b + \epsilon_t + \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \ldots + \alpha_{t-q} \epsilon_{t-q}$$

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = a_1 Y_{t-2} + a_2 Y_{t-3} + \ldots + a_0 Y_0 + \epsilon_{t-1}$$

ARIMA

- O modelo ARIMA combina os componentes AR e MA, assumindo que a série temporal foi diferenciada pelo menos uma vez para ficar estacionária.
- Log, teremos o modelo ARIMA a partir de AR e MA :

$$Y_t = b + a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_p Y_{t-p} + \\ + \epsilon_t + \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \ldots + \alpha_{t-q} \epsilon_{t-q}$$