

A1: SÉRIES TEMPORAIS

Introdução

Uma série temporal é um conjunto de observações y_t registradas num tempo t específico, $\{y_t\}_{t=1}^T$. Um modelo para $\{y_t\}_{t=1}^T$ é uma especificação da distribuição conjunta de uma sequência de variáveis aleatórias $\{Y_t\}_{t=1}^T$, da qual esperamos que $\{y_t\}_{t=1}^T$ seja uma realização.

Na prática utilizamos primeiro e segundo momentos para aproximar através de uma distribuição normal. O **modelo linear padrão** é dado por: $y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_p x_{p,t} + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Matricialmente, temos $Y_{t \times 1} = X_{t \times p} \beta_{p \times 1} + \epsilon_{t \times 1}$, de modo que $Y|X \sim \mathcal{N}(X\beta, \sigma^2 \mathbb{I})$.

Para minimizar o erro quadrático (OLS, o MLE do modelo linear padrão) basta fazermos $\hat{\beta} = (X^T X)^{-1} X^T Y$. O erro esperado $\mathbb{E}[(y_t - \hat{y}_t)^2]$ nos leva ao dilema viés-variância. Modelos muito simples tendem a ter viés grande e variância baixa; modelos muito complexos tendem a ter viés baixo e variância alta.

$$\text{Bias}[\hat{y}_t] = \mathbb{E}[\hat{y}_t - y_t]$$

$$\text{Var}[\hat{y}_t] = \mathbb{E}[(\hat{y}_t)^2] - (\mathbb{E}[\hat{y}_t])^2$$

$$\mathbb{E}[(y_t - \hat{y}_t)^2] = \text{Bias}^2 + \text{Variance} + \text{noise}$$

Sendo T o número de observações e P o número de covariáveis, considere:

(1) $T \gg P$: baixa variância

(2) $T \approx P$: alta variância

(3) $T < P$: variância infinita

Para solucionar os casos (2) e (3), normalmente utilizamos métodos de regularização como Lasso (L1) e Ridge (L2):

$$\text{Lasso: } \sum_{t=1}^T (y_t - \hat{y}_t)^2 + \lambda \sum_{j=1}^P |\beta_j|$$

$$\text{Ridge: } \sum_{t=1}^T (y_t - \hat{y}_t)^2 + \lambda \sum_{j=1}^P \beta_j^2$$

Caso a relação entre $\{y_t\}$ e X não seja linear, teremos viés alto aplicando modelos lineares. Podemos utilizar modelos não lineares como Generalized Additive Models ou modelos de Deep Learning.

Os modelos de série temporal são importantes quando existe uma dependência temporal entre os resíduos que não foi removida pelos modelos anteriores. Nesse caso, podemos modelar a dependência para entender melhor previsões sobre o futuro.

Definições

Para identificar a dependência nos resíduos, precisamos de algumas definições

→ Seja $\{Y_t\}$ um modelo de séries temporais com $\mathbb{E}(Y_t^2) < \infty$:

A função média de $\{Y_t\}$ é $\mu_Y(t) = \mathbb{E}[Y_t]$.

A função de autocovariância de $\{Y_t\}$ é $\gamma_Y(r, s) = \text{cov}(X_r, X_s) = \mathbb{E}[(X_r - \mu_Y(r))(X_s - \mu_Y(s))]$.

→ $\{Y_t\}$ é fracamente estacionária se:

(i) $\mu_Y(t)$ é independente de t

(ii) $\gamma_Y(t+h, t)$ é independente de t para cada t .

→ Se $\{Y_t\}$ é uma série temporal estacionária:

(i) a função de autocovariância (ACVF) de $\{Y_t\}$ no lag h é $\gamma_Y(h) = \text{cov}(X_{t+h}, X_t)$

(ii) a função de autocorrelação (ACF) no lag h é $f_Y(h) = \frac{\gamma_Y(h)}{\gamma_Y(0)} = \text{cov}(X_{t+h}, X_t)$.

Por exemplo: No caso de uma série temporal $\{Y_t\} \sim \text{IID}(0, \sigma^2)$, variância finita. A média é $\mathbb{E}[Y_t] = 0 \forall t$ e $\gamma_Y(t+h, h) = \sigma^2$ se $h = 0$; 0 se $h \neq 0$. Trata-se de um caso particular de série fracamente estacionária, já que a média é constante e a covariância depende apenas da distância entre os tempos.

OBS.: se a série segue distribuição White Noise, $\{Y_t\} \sim \text{WN}(0, \sigma^2)$, as variáveis aleatórias são não correlacionadas, embora não sejam necessariamente independentes. Ou seja, toda sequência IID é WN, mas nem toda sequência WN é IID.

→ Seja y_1, \dots, y_T observações de uma série temporal:

$$\text{Média amostral: } \bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$$

Função de autocovariância amostral: $\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-|h|} (y_{t+|h|} - \bar{y})(y_t - \bar{y})$

Função de autocorrelação amostral: $\hat{f}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, -T < h < T$.

→ Se a série temporal tem distribuição IID($0, \sigma^2$) e variância finita, podemos mostrar que a função de autocorrelação é aproximadamente IID($0, \frac{1}{T}$) para T grande. Por isso, esperamos que 95% das autocorrelações amostrais deve estar entre os intervalos $\pm \frac{1.96}{\sqrt{T}}$. Assim, podemos observar os resíduos e identificar se há relação temporal ou não.

O estimador linear só é ótimo no caso Gaussiano

Já sabemos como identificar dependência temporal através de $\hat{f}(h)$. Vamos entender o papel dessa dependência na previsão de valores futuros. Para uma série temporal Gaussiana $\{Y_t\}$, queremos prever $Y_{n+h}|Y_n = y_n$. Sabemos das propriedades das Gaussianas que $Y_{n+h}|Y_n = y_n \sim \mathcal{N}[\mu + f(h)(y_n - \mu), \sigma^2(1 - f(h)^2)]$.

Seja $m(Y_n)$ uma estimativa pontual. Minimizar o erro quadrático dessa estimativa é dada por $\text{MSE} = \mathbb{E}[Y_{n+h} - m(Y_n)]^2 = \sigma^2(1 - f(h)^2)$. Observe que, se não assumirmos normalidade, minimizar o MSE se torna mais complicado, geralmente utilizando algum método computacional.

Em outras palavras, o estimador linear $l(Y_t) = aY_t + b$ da média só é ótimo no caso Gaussiano, onde $l(Y_t) = m(Y_n)$. Funções lineares funcionam bem no mundo Gaussiano, assim como primeiro e segundo momentos são suficientes. Então, assumimos que a série temporal é Gaussiana e usamos a média e a variância para que o modelo seja útil de alguma forma e mais fácil de ser computado.

Modelagem de dependência temporal

Todo modelo estacionário é um processo linear ou pode ser transformado em um processo linear ao subtrair uma componente determinística.

Motivação: observamos $\{Y_t\} = y_1, y_2, \dots, y_T$, sem conhecer o processo gerador dos dados. Desejamos obter estimativas h passos a frente.

Estimativas baseadas em modelos

Método da Média: $\hat{y}_{T+h|T} = \hat{y} = \frac{y_1 + \dots + y_T}{T}$

Naive ou Random Walk with no Drift:

$$y_t = y_{t-1} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\hat{y}_{T+h|T} = y_t$$

Seasonal Naive: $\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$, onde m é o período da sazonalidade e $k = \lfloor (h-1)/m \rfloor$ (arredonda pra baixo).

Para dados mensais, a previsão para o próximo mês de fevereiro é igual à última observação de fevereiro. Para dados trimestrais, previsão para o primeiro trimestre é igual ao do último trimestre observado.

Random Walk with Drift: detecção de tendências.

$$y_t = c + y_{t-1} + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

$$\hat{y}_{T+h|T} = y_T + h \cdot c$$

$$c = \frac{y_T - y_1}{T - 1}$$

$$\hat{y}_{T+h|T} = y_T + h \cdot \frac{y_T - y_1}{T - 1}$$

Valores ajustados

Valores ajustados quase sempre envolvem uma predição um passo a frente, i.e., $\hat{y}_{t|t-1}, y_t|y_1, \dots, y_{t-1}$. Embora sejam utilizados para analisar o modelo, não podem ser utilizados para medir a capacidade de generalização – afinal, todos os parâmetros foram estimados utilizando todos os pontos de dado disponíveis.

Resíduos: $\epsilon_t = y_t - \hat{y}_t$, i.e., o que sobra depois de ajustado o modelo. Transformações nos dados do tipo $w_t = \log(y_t)$ nos levam a calcular e analisar os resíduos transformados por $\epsilon_t = w_t + \hat{w}_t$ (resíduos de inovação).

Diagnósticos sobre os resíduos

Bons modelos de previsão terão:

- resíduos não correlacionados. Caso estejam correlacionados, alguma informação não foi capturada pelo modelo;
- resíduos com média zero. Caso contrário, as previsões estão viesadas.

São boas características (para facilitar o cálculo de intervalos de confiança), embora não sejam estritamente necessárias:

- variância constante nos resíduos (homocedasticidade)
- resíduos normalmente distribuídos

Distribuição das previsões

Na maioria das vezes, assumimos que modelos são Gaussianos, então a previsão pontual é a média da distribuição Gaussiana. Portanto, supondo que queremos calcular um intervalo de confiança de 95%, temos: $\hat{y}_{T+h|T} \pm 1.96\sqrt{\text{Var}(\hat{y}_{T+h|T})}$.

Intervalos de confiança one-step: podem ser estimados a partir do desvio padrão dos resíduos:

$$\hat{\sigma} = \sqrt{\frac{1}{T - K - M} \cdot \sum_{t=1}^T \epsilon_t^2},$$

onde K é o número de parâmetros estimados e M é o número de resíduos perdidos (faltantes).

Ou seja, começamos estimando o desvio padrão um passo a frente e depois calculamos o desvio padrão h passos a frente:

$$\textbf{Média: } \hat{\sigma}_h = \hat{\sigma} \sqrt{1 + \frac{1}{T}}$$

$$\textbf{Naive: } \hat{\sigma}_h = \hat{\sigma} \sqrt{h}$$

$$\textbf{Seasonal Naive: } \hat{\sigma}_h = \hat{\sigma} \sqrt{k+1}$$

$$\textbf{Drift: } \hat{\sigma}_h = \hat{\sigma} \sqrt{h \left(\frac{1+h}{T-1} \right)}$$

E se não for razoável assumir normalidade?

Nesse caso, podemos usar bootstrapping para estimar os intervalos de confiança. Assumindo que os resíduos são independentes e que a variância é constante, fazemos: $\epsilon_t = y_t - \hat{y}_{t|t-1} \Rightarrow y_t = \hat{y}_{t|t-1} + \epsilon_t$. Com isso conseguimos gerar novas observações $y_{t+1} = \hat{y}_{t+1|t} + \epsilon_{t+1}$, assumindo que os erros futuros sejam semelhantes aos erros passados. Fazendo o processo repetidamente, conseguimos gerar uma distribuição de previsões futuras para calcular estimativas.

Bootstrap é uma ferramenta para calcular incerteza, mas, se você tem poucos dados, a incerteza já é grande e será refletida no bootstrap.

Transformação Box-Cox: se os dados apresentam variações que crescem ou diminuem dependendo do nível da série, ou seja, não respeita homocedasticidade, utilizamos a transformação Box-Cox para estabilizar a variância:

$$w_t = \begin{cases} \frac{y_t^\lambda - 1}{\lambda}, & \text{se } \lambda > 0 \\ \log(y_t), & \text{se } \lambda = 0 \end{cases}$$

O parâmetro λ é estimado a partir dos dados, como um hiper-parâmetro. Quando $\lambda = 0 \Rightarrow w_t = \log(y_t) \Rightarrow y_t = \exp(w_t)$.

$$\mathbb{P}(w_1 \leq w_t \leq w_2) = 0.95$$

$$\mathbb{P}(w_1 \leq \log(y_t) \leq w_2) = 0.95$$

$$\mathbb{P}(\exp(w_1) \leq y_t \leq \exp(w_2)) = 0.95$$

Decomposição de Séries Temporais

Uma série temporal é composta por $y_t = S_t + T_t + R_t$, onde S_t é o componente sazonal, T_t é o componente de tendência e R_t é o componente restante.

Passo 1: Estimar a tendência. Sendo d o período da sazonalidade e q o número de pontos à esquerda e à direita do ponto t , podemos estimar T_t como:

$$T_t = \hat{m}_t = \frac{0.5x_{t-q} + x_{t-q+1} + \dots + x_{t+q-1} + 0.5x_{t+q}}{d}$$

Caso d não seja par, não conseguimos pegar o período de maneira simétrica. Para não deixar ninguém de fora, damos metade do peso para cada uma das pontas. Num caso com $d = 4$, por exemplo, $q = 2$ e $T_t = \frac{0.5x_{t-2} + x_{t-1} + x_{t+1} + 0.5x_{t+2}}{4}$.

Passo 2: Calcular o componente sazonal.

Removemos a tendência para calcular a sazonalidade. Para cada $k = 1, 2, \dots, d$, calculamos a média dos desvios.

$$w_k = \{(x_{k+jd} - \hat{m}_{k+jd})\}, \quad q < k + jd < n - q,$$

onde j é o número de períodos completos (o elemento “circular”), k é o número de pontos dentro de um período, w_k é a média dos pontos dentro de um período e d é o período da sazonalidade.

É importante tirar a média dos valores de w . No caso abaixo, por exemplo, tiraríamos a média dos dois valores de w_1 :

$$\begin{aligned} w_1 &= x_{1+0 \times 12} = x_1 - m_1 \\ &= x_{1+1 \times 12} = x_{13} - m_{13} \end{aligned}$$

Para garantir que a média da componente sazonal é zero, fazemos:

$$\hat{S}_k = w_k - \frac{1}{d} \sum_{i=1}^d w_i, \quad k = 1, \dots, d$$
$$\hat{S}_k = \hat{S}_{k-d}, \quad k > d$$

Com isso podemos remover o componente sazonal da série com $d_t = x_t - \hat{S}_t$ e fazer previsões a partir da decomposição $y_t = \hat{S}_t + \hat{A}_t$, onde $\hat{A}_t = T_t + R_t$ (a componente sem sazonalidade).

Avaliação de previsões pontuais

- Importante olhar os resíduos
- Importante olhar predições fora da amostra (evitar overfitting)

Erros de previsão

A partir de $\hat{y}_{T+h|T}$, uma previsão fora da amostra, podemos calcular o resíduo fora da amostra $\epsilon_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T}$.

Considere as métricas Mean Absolute Error (MAE) e Root Mean Squared Error (RMSE), que por natureza estão na mesma escala dos dados:

MAE: $\text{mean}(|\epsilon_t|)$

RMSE: $\sqrt{\text{mean}(\epsilon_t^2)}$

Erros percentuais têm a vantagem da escala “padronizada”, por exemplo **MAPE:** $P_t = \frac{100 \cdot \epsilon_t}{y_t}$. Essa no entanto, é uma métrica que não pode ter valores iguais a 0 e é instável para valores próximos de 0. Nesse caso, usamos:

$$q_j = \frac{\epsilon_j}{\frac{1}{T-m} \sum_{t=1}^T |y_t - y_{t-m}|} \text{ ou } \frac{\epsilon_j}{\frac{1}{T-m} \sum_{t=1}^T (y_t - y_{t-m})^2},$$

fazemos $\text{MASE} = \text{mean}(|q_j|)$ e $\text{RMSSE} = \sqrt{\text{mean}(q_j^2)}$. Se a série temporal não for sazonal, basta definir $m = 1$.

Avaliação das distribuições das previsões

Quando usamos medidas como RMSE ou MAE, estamos olhando para a média e mediana da distribuição, respectivamente. Há métricas que nos permitem avaliar quaisquer quantis, não apenas o 0.5 (mediana). Para isso, usamos o Quantile Score:

$$Q_{p,t} = \begin{cases} 2(1-p)(f_{p,t} - y_t), & \text{se } y_t < f_{p,t} \\ 2(p)(y_t - f_{p,t}), & \text{se } y_t \geq f_{p,t} \end{cases}$$

O modelo mais calibrado é aquele que tem o menor quantile score para aquele quantil, o que significa que a sua distribuição empírica está caindo mais próxima da distribuição teórica.

Outra medida é o Winkler Score, para intervalos de confiança. Intervalo de confiança de $100(1-\alpha)\%$ no tempo t : $[l_{\alpha,t}, u_{\alpha,t}]$:

$$W_{\alpha,t} = \begin{cases} (u_{\alpha,t} - l_{\alpha,t}) + \frac{2}{\alpha}(l_{\alpha,t} - y_t), & \text{se } y_t < l_{\alpha,t} \\ (u_{\alpha,t} - l_{\alpha,t}), & \text{se } l_{\alpha,t} \leq y_t \leq u_{\alpha,t} \\ (u_{\alpha,t} - l_{\alpha,t}) + \frac{2}{\alpha}(y_t - u_{\alpha,t}), & \text{se } y_t > u_{\alpha,t} \end{cases}$$

$$l_{\alpha,t} = f_{\frac{\alpha}{2},t} \quad u_{\alpha,t} = f_{1-\frac{\alpha}{2},t}$$
$$W_{\alpha,t} = \frac{Q_{\frac{\alpha}{2},t}, Q_{1-\frac{\alpha}{2},t}}{\alpha}$$

Para avaliar toda a distribuição, temos o Continuous Ranked Probability Score (CRPS):

$$\text{CRPS}(F, y_t) = \int_{-\infty}^{\infty} (\underbrace{F(x)}_{\mathbb{P}(Y_t \leq x)} - 1(x - y_t))^2 dx$$
$$1(x - y_t) = \begin{cases} 1, & \text{se } x \geq y_t \\ 0, & \text{se } x < y_t \end{cases}$$

Modelos de Regressão para Séries Temporais

Modelos de regressão impõem algumas hipóteses:

- Espera-se que a relação entre y e x seja linear;
- Espera-se que os resíduos tenham média zero, não sejam auto-correlacionados e não sejam correlacionados com x . É bom, mas não necessário, que sejam normalmente distribuídos e apresentem homocedasticidade.
- Embora seja uma hipótese que não é verdadeira na maioria dos casos reais, espera-se que x não seja variável aleatória. Normalmente não conhecemos x perfeitamente e, por isso, ele é fonte de incerteza, mas essa incerteza não é levada em conta no cálculo de intervalos de confiança, por exemplo.

Sendo $y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t} + \epsilon_t$:

Método dos Mínimos Quadrados: $\min_{\beta's} \sum (y_t - \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t})^2$, obtendo estimativas para os parâmetros $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \dots + \hat{\beta}_k x_{k,t}$.

Coefficiente de Determinação (R^2): $1 - \frac{\sum (\hat{y}_t - \bar{y})^2}{\sum (y_t - \bar{y})^2}$. Trata-se de uma medida de ajuste de modelo (e, portanto, depende do treino) – quanto mais próximo de 1, melhor o ajuste; quanto mais próximo de 0, pior o ajuste. Não podemos nos orientar somente por essa métrica, porque, por exemplo, a adição de variáveis sempre a torna mais próxima de 1 e bonifica o overfitting. Em outras palavras, não é capaz de medir generalização do modelo.

Coefficiente de Determinação Ajustado (R_{adj}^2):

$1 - \frac{\text{RSS}/(T - k - 1)}{\text{TSS}/(T - 1)}$. Tenta corrigir para avaliar também a capacidade de generalização do modelo.

Desvio padrão da regressão¹: $\hat{\sigma} = \sqrt{\frac{1}{T - K - 1} \sum_{t=1}^T \epsilon_t^2}$

Verificação de tendência e sazonalidade:

$$y_t = \underbrace{\beta_0 + \beta_1 t}_{\text{tendência}} + \underbrace{\beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t}}_{\text{trimestral}} + \epsilon_t$$

Se sabemos $\beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t}$, sabemos d_1 , que é o trimestre que não está ali e é dado por $d_1 = 1 - d_2 - d_3 - d_4$. Se o adicionássemos, não teríamos solução, porque a matriz de design não teria inversa.

Formulação matricial

$$\underline{y} = (y_1, \dots, y_T)^T, \underline{\epsilon} = (\epsilon_1, \dots, \epsilon_T)^T, \underline{\beta} = (\beta_1, \dots, \beta_k)^T$$
$$X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,T} & \dots & x_{k,T} \end{pmatrix}$$

¹Graus de liberdade: $T - K - 1$. Na prática, grau de liberdade significa que não conseguimos, livremente, escolher os valores de todos os parâmetros. Quando fazemos o ajuste de um modelo linear com dois parâmetros e 10 valores, por exemplo, temos 8 graus de liberdade.

Pela estimação de mínimos quadrados, $\hat{\beta} = (X^T X)^{-1} X^T y$. A matriz $X^T X$ precisa ser invertível, de modo que X precisa ser posto completo e nenhuma coluna pode ser uma combinação linear das outras (como vimos no exemplo).

$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$, conhecida como Hat Matrix.

LOOCV: $\frac{1}{T} \sum_{t=1}^T \left(\frac{y_t - \hat{y}_t}{1 - h_{t,t}} \right)$. Para cada ponto, ajustamos o modelo sem aquele ponto e calculamos o erro. Fazemos isso para todos os pontos e calculamos a média. É um método de validação cruzada. Mas, usando a Hat Matrix, podemos fazer isso de forma mais eficiente ajustando o modelo apenas uma vez:

Sendo x^* novos dados, considere:

Previsões: $\hat{y} = x^* + \hat{\beta} = x^*(X^X)^{-1} X^T y$

Variância das previsões: $\hat{\sigma}_\epsilon^2 \cdot [1 + x^*(X^T X)^{-1} x^{*T}]$

Intervalo de confiança das previsões:

$\hat{y} \pm 1.96 \cdot \hat{\sigma}_\epsilon \cdot \sqrt{1 + x^*(X^T X)^{-1} x^{*T}}$. O intervalo de confiança leva em consideração a incerteza relativa ao termo ϵ e à estimação dos coeficientes, mas não leva em consideração os erros em x^* .

Exponential Smoothing

Nos modelos estudados que tratam da modelagem de dependência temporal, há casos em que todo o peso para a previsão futura vai para a última observação (Naive Method) ou todas as observações possuem o mesmo peso no cálculo da próxima previsão (Average Method).

Exponential Smoothig simples: os pesos decaem exponencialmente à medida em que as observações ficam mais distantes.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1 - \alpha)y_{T-1} + \alpha(1 - \alpha)^2 y_{T-2} + \dots$$

Na forma de média ponderada, temos:

$$\begin{aligned} \hat{y}_{T+1|T} &= \alpha y_T + (1 - \alpha)\hat{y}_{T|T-1} \\ &\vdots \\ \hat{y}_{T+1|T} &= \sum_{j=0}^{T-1} \alpha(1 - \alpha)^j y_{T-j} + (1 - \alpha)^T l_0 \end{aligned}$$

Precisamos estimar α, l_0 .

Flat forecast function: $\hat{y}_{T+h|T} = l_T$. Útil apenas se os dados não possuem tendência e sazonalidade.

⇒ **State space model para exponential smoothing simples:**

$$\rightarrow \hat{y}_{T+1|T} = \alpha y_T + (1 - \alpha)\hat{y}_{T|T-1}$$

$$\rightarrow \text{Definimos } \hat{y}_{T+1|T} = l_T$$

$$\text{Equação de forecast: } \hat{y}_{T+1|T} = l_T$$

$$\text{Equação de smoothing: } l_T = \alpha y_T + (1 - \alpha)l_{T-1}$$

$$\begin{aligned} l_t &= \alpha y_T + l_{T-1} - \alpha l_{T-1} \\ l_t &= l_{T-1} + \alpha(y_T - l_{T-1}) \\ l_T &= l_{T-1} + \alpha(y_T - \hat{y}_{T|T-1}) \\ l_T &= l_{T-1} + \alpha\epsilon_T \end{aligned}$$

Ou seja, o erro dos dados de treino leva ao ajuste do nível estimado, e α controla a correção.

$$y_t = l_{T-1} + \epsilon_T, \epsilon_T \sim \mathcal{N}(0, \sigma^2)$$

$$l_T = l_{T-1} + \alpha\epsilon_T$$

⇒ **Método de tendência linear de Holt:**

$$\text{Equação de forecast: } \hat{y}_{T+h|T} = l_{T+h} + b_T$$

$$\text{Equação de nível: } l_T = \alpha y_T + (1 - \alpha)(l_{T-1} + b_{T-1})$$

$$\text{Equação de tendência: } b_T = \beta^*(l_T - l_{T-1}) + (1 - \beta^*)b_{T-1}$$

Formulação do modelo de state-space:

$$\begin{aligned} y_T &= l_{T-1} + b_{T-1} + \epsilon_T \\ l_T &= l_{T-1} + b_{T-1} + \alpha\epsilon_T \\ b_T &= b_{T-1} + \alpha\beta^*\epsilon_T \end{aligned}$$

⇒ **Damped Trend Methods:** projects the slope into the future indefinitely.

$$\begin{aligned} y_{T+h} &= l_T + (\phi + \phi^2 + \dots + \phi^h)b_T \\ l_T &= \alpha y_T + (1 - \alpha)(l_{T-1} + \phi b_{T-1}) \\ b_T &= \beta^*(l_T - l_{T-1}) + (1 - \beta^*)\phi b_{T-1} \end{aligned}$$

Se $\phi = 1$, esse é o método de Holt.

$$y_{T+h|T} = l_T + \frac{\phi b_T}{1 - \phi} \text{ à medida que } h \rightarrow \infty.$$

⇒ **Holt-Winters seasonality method:**

$$\begin{aligned} y_{T+h} &= l_T + h b_T + S_{t+h-m(k+1)} \\ l_T &= \alpha(y_T - S_{T-m}) + (1 - \alpha)(l_{T-1} + b_{T-1}) \\ b_T &= \beta^*(l_T - l_{T-1}) + (1 - \beta^*)b_{T-1} \\ S_T &= \gamma^*(y_T - l_{T-1} - b_{T-1}) + (1 - \gamma^*)S_{T-m} \end{aligned}$$

State-space representation:

$$\begin{aligned} y_T &= l_{T-1} + b_{T-1} + S_{t+m}) + \epsilon_T \\ l_T &= \alpha l_{T-1} + b_{T-1} + \alpha\epsilon_T \\ b_T &= b_{T-1} + \beta\epsilon_T \\ S_T &= S_{T-m} + \gamma\epsilon_T \end{aligned}$$

⇒ **Holt-Winters seasonality with multiplicative method:** ideal quando variações sazonais mudam de maneira proporcional ao nível da série.

$$\begin{aligned} \hat{y}_{T+h} &= (l_T + h b_T) S_{t_h - m(k+1)} \\ l_T &= \alpha \frac{y_T}{S_{t-m}} + (1 + \alpha)(l_{T-1} + b_{T-1}) \\ b_T &= \beta^*(l_T - l_{T-1}) + (1 - \beta^*)b_{T-1} \\ S_T &= \gamma^* \gamma \frac{y_T}{l_{T-1} + b_{T-1}} + (1 - \gamma)S_{T-m} \end{aligned}$$