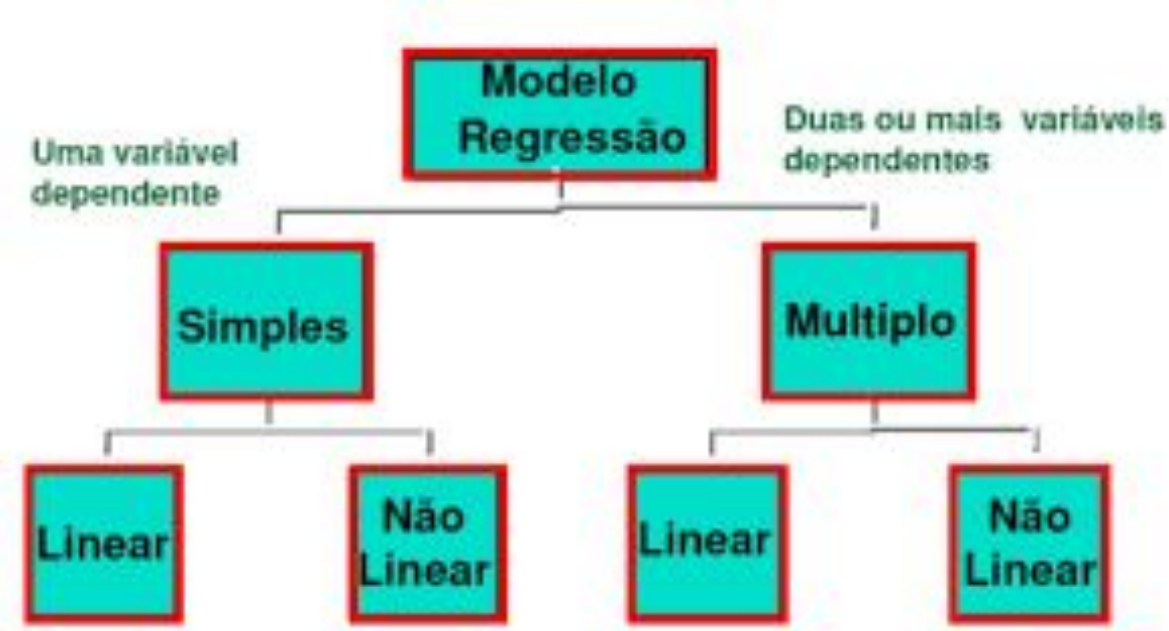
Regressão Linear Múltipla

Regressão


Modelo Regressão
Uma variável dependente
Duas ou mais variáveis dependentes
Simples
Multiplo
Linear
Não Linear
Linear
Não Linear

Regressão Linear Múltipla

- Temos uma variável dependente e duas ou mais variáveis independentes
- Tentamos explicar linearmente a variação da VD com as nossas VIs
- Inferência e Predição são tarefas que podem ser realizadas

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j + E$$

Estimação dos parâmetros

- É necessário estimar os coeficientes, já que não são conhecidos

- Método dos mínimos quadrados: Tentamos minimizar a soma dos erros quadráticos do modelo e com isso chegamos a estimativas dos nosso coeficientes.

- Assim, após estimarmos nossos coeficientes precisamos testar a validade deles.

- Para estimarmos os parâmetros precisariamos realizar multiplicações de matrizes, inversões, transposições, etc...

Métricas de Validação

- Visam verificar o quão satisfatoriamente a equação de regressão estimada ajusta os dados.
- R2 -> Coeficiente de determinação
- Interpretação: Porcentagem da soma total dos quadrados que pode ser explicada usando a reta de regressão. Ou seja, se R2 = 0.9, então 90% da variabilidade da VD pode ser explicada por meio da relação linear com as VIs.

Sum of Squares Total →

$$SST = \sum (y - \bar{y})^2$$

Sum of Squares Regression →

$$SSR = \sum (y' - \bar{y'})^2$$

Sum of Squares Error →

$$SSE = \sum (y - y')^2$$

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

Problemas com o R2

- Regressão Simples -> R2 era usado para explicar a variabilidade de Y em relação a X
- Entretanto, na Reg. Múltipla temos diversas covariáveis
- R2-Score -> Usado para comparar modelos com a mesma quantidade de covariáveis
- Modelos com diferentes nros de variáveis não podem ser comparados usando o R2;
- R2 irá aumentar para modelos com mais variáveis, podendo gerar overfitting

Coeficiente de Determinação ajustado

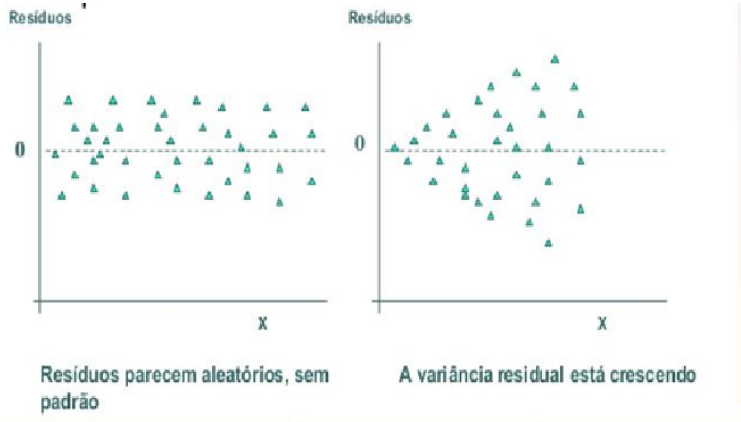
- R2 Ajustado leva em consideração o nro de variáveis independentes no modelo
- Penaliza o aumento do R2 que ocorre quando mais VI são adicionadas, evitando overfitting
- Quanto mais VIs, maior será a penalização

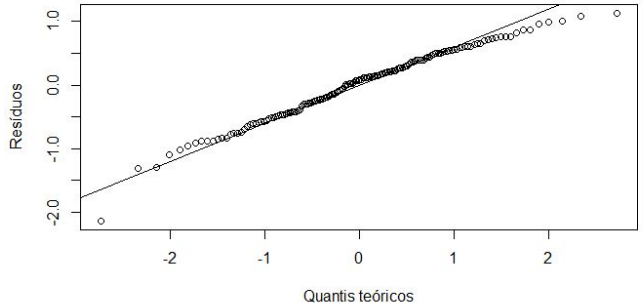
$$R^2_{ajust} = 1 - (1 - R^2)(\frac{n-1}{n-q-1})$$

N -> Nro de amostras
q -> Nro de VIs

Suposições do modelo

- Para realizarmos a análise da regressão é necessário fazer algumas suposições sobre o nosso modelo.
  1. O erro é uma variável aleatória com valor esperado igual à 0
  2. Variância do erro é a mesma para todos os valores da variável independente. (Podemos verificar em um grafo que contenha Var(Erro) x VIs ou VD)
  3. Erro é independente
  4. O erro é normalmente distribuído (Podemos verificar com QQplot)





Teste de Significância

- Utilizado para testar se uma relação de regressão é significativa
- No teste, iremos testar se os coeficientes são iguais a zero ou diferentes de zero
- Como os coeficientes foram estimados anteriormente, precisamos verificar se eles podem ser utilizados
- O teste tem o seguinte formato: (H0: Hip. Nula // H1: Hip. Alternativa)

$$IGC_i = \beta_0 + \beta_1 fastfood_i + u_i$$

$$\begin{cases} H_0: \beta_1 = 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ não é relevante para explicar o IGC de um indivíduo} \\ H_A: \beta_1 \neq 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ é relevante para explicar o IGC de um indivíduo} \end{cases}$$

Teste-F

- Teste-F: verificar se o modelo é significativo, ou seja, verificar se pelo menos um dos coeficientes angulares deve ser aceito.
- Assim, utilizamos a estatística F, que segue uma distribuição de **Fisher-Snedecor**, com **q e n-q-1** graus de liberdade.
- Testar a significância geral do modelo global.

$$H_0 : \beta_1 = 0 \& \beta_2 = 0 \& \beta_3 = 0 \& \ldots \beta_q = 0$$

$$H_1 : \beta_1 \neq 0 | \beta_2 \neq 0 | \beta_3 \neq 0 | \ldots \beta_q \neq 0$$

Interpretação

- Ao aceitarmos a hipótese nula, entendemos que todos os coeficientes são igual a zero, e dessa forma não podermos explicar Y linearmente pelas cováriaveis do nosso conjunto.
- Caso a hipótese nula fosse recusada, pelo menos um dos coeficientes do nosso modelo é diferente de zero
- Caso a hipótese nula seja recusada, significa que temos pelo menos um coeficiente que é significativo para explicar Y linearmente,
- Para verificar qual/quais coeficientes são significativos aplicamos o teste-t individualmente, so que ao invés de n-2 graus de liberdade, teremos n-q-1 graus de liberade.

Exemplo prático

OLS Regression Results

Dep. Variable:	Lottery	R-squared:	0.338
Model:	OLS	Adj. R-squared:	0.287
Method:	Least Squares	F-statistic:	6.636
Date:	Sat, 28 Nov 2020	Prob (F-statistic):	1.07e-05
Time:	14:39:43	Log-Likelihood:	-375.30
No. Observations:	85	AIC:	764.6
Df Residuals:	78	BIC:	781.7
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656

Omnibus:	3.049	Durbin-Watson:	1.785
Prob(Omnibus):	0.218	Jarque-Bera (JB):	2.694
Skew:	-0.340	Prob(JB):	0.260
Kurtosis:	2.454	Cond. No.	371.

Predição

- Predição é uma tarefa diferente da inferência, e dessa maneira, podemos abdicar de algumas formalidades existentes
- Predição **não precisa necessariamente de teste de hipóteses para validar um modelo**
- Predição **não precisa que o modelo tenha suposições**
- Predição **não precisa verificar multiColinearidade**
- Predição precisa de : Métricas de validação bem comportadas, Controle do overfitting e seleção de variáveis (feature selection)
- Lembrar que para comparar modelos com diferentes nros de VIs -> R2 Ajustado