Análise de Suposições do Modelo Linear

ANA CAROLINA

Graduada e Mestre em Estatística

- Fundadora do Descomplica Estatística
- Cientista de Dados em TakeBlip

AGENDA

- **Suposições do modelo**
  Teste de independência e normalidade

- **Suposições do modelo**
  Linearidade, homocedasticidade, multicolinearidade e pontos discrepantes

- **Alternativas para a violação dos pressupostos da regressão linear**

Suposições do modelo

Testes de independência e normalidade dos resíduos

Regressão linear simples e múltipla

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p + \varepsilon$$

y = Variável dependente/ target/ alvo /resposta

x = Variáveis independentes/explicativas/ features

ε = Erro (valor real - valor estimado pelo modelo)

Resíduos

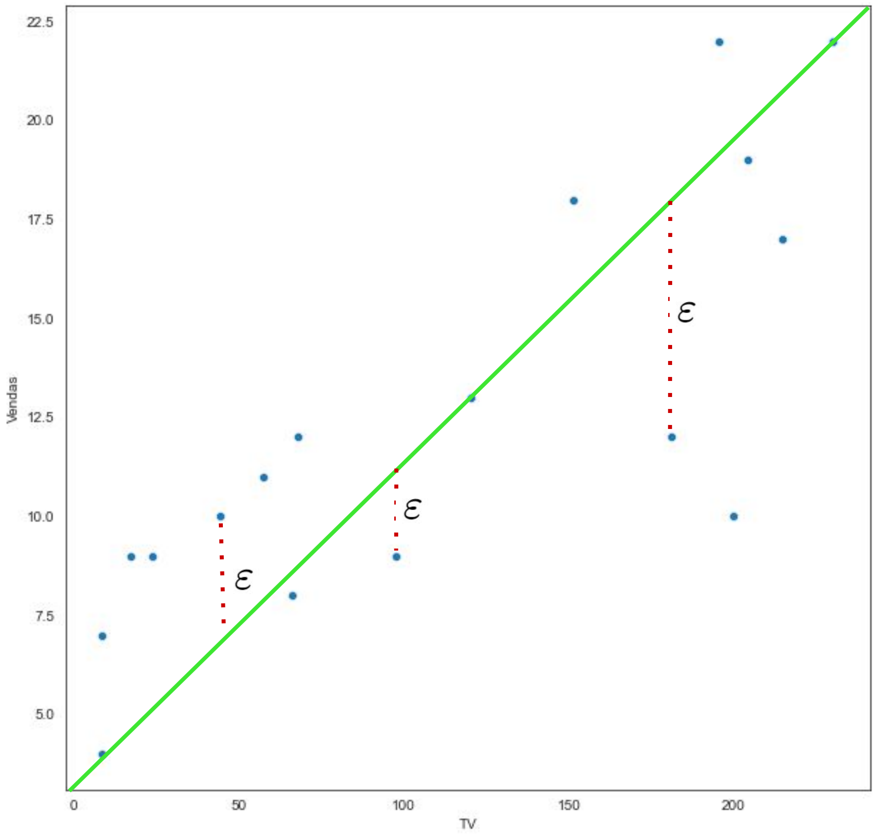
$$\hat{y} = \hat{\beta}_0 + \hat{\beta} X$$

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

Valor que queremos estimar

Função matemática que vamos usar para fazer a estimação

- **Representa a quantidade da variabilidade de Y que o modelo ajustado não consegue explicar.**



Como garantir que o meu modelo é bom?

- Verificar a qualidade de ajuste do modelo.
- Avaliar se as suposições do modelo são atendidas (através dos resíduos).
  - Os resíduos contém informação sobre o motivo do modelo não ter se ajustado bem aos dados. Eles conseguem indicar se uma ou mais suposições do modelo foram violadas.

Verificando a qualidade de ajuste - R^2

- **Regressão linear múltipla**

```python
# Construção do modelo

import statsmodels.formula.api as smf

modelo_ls1 = smf.ols(formula = 'Fatura ~ Limite', data=credit).fit()
modelo_ls1.summary()
```

OLS Regression Results

Dep. Variable:	Fatura	R-squared:	0.633
Model:	OLS	Adj. R-squared:	0.632
Method:	Least Squares	F-statistic:	531.8
Date:	Sun, 03 Jul 2022	Prob (F-statistic):	4.71e-69
Time:	15:53:23	Log-Likelihood:	-2151.8
No. Observations:	310	AIC:	4308.
Df Residuals:	308	BIC:	4315.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-209.2924	40.750	-5.136	0.000	-289.476	-129.109
Limite	0.1605	0.007	23.060	0.000	0.147	0.174

Omnibus:	8.813	Durbin-Watson:	1.903
Prob(Omnibus):	0.012	Jarque-Bera (JB):	9.619
Skew:	0.313	Prob(JB):	0.00815
Kurtosis:	3.595	Cond. No.	1.67e+04

Proporção da variação total de Y que está sendo explicada pelo modelo de regressão (entre 0 e 1) .

Quanto mais próximo de 1, mais o modelo explica.

- **Regressão linear múltipla**

```python
modelo_ls = smf.ols('Fatura ~ Limite  + Idade + Estudante', data = credit).fit()
modelo_ls.summary()
```

OLS Regression Results

Dep. Variable:	Fatura	R-squared:	0.747
Model:	OLS	Adj. R-squared:	0.745
Method:	Least Squares	F-statistic:	301.7
Date:	Sun, 03 Jul 2022	Prob (F-statistic):	4.74e-91
Time:	16:08:15	Log-Likelihood:	-2094.1
No. Observations:	310	AIC:	4196.
Df Residuals:	306	BIC:	4211.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-160.8979	48.390	-3.325	0.001	-256.117	-65.679
Estudante[T.Yes]	394.7980	36.106	10.934	0.000	323.750	465.846
Limite	0.1728	0.006	29.182	0.000	0.161	0.184
Idade	-2.9817	0.695	-4.288	0.000	-4.350	-1.614

Omnibus:	6.223	Durbin-Watson:	1.978
Prob(Omnibus):	0.045	Jarque-Bera (JB):	6.393
Skew:	-0.340	Prob(JB):	0.0409
Kurtosis:	2.818	Cond. No.	2.43e+04

Medidas de qualidade do ajuste

Teste F

- **Hipótese nula**: O ajuste do modelo somente com o intercepto e o do seu modelo são iguais.

$$y = \beta_0 + \varepsilon \qquad = \beta_0 + \beta_1 x + \varepsilon$$

- **Hipótese alternativa:** O ajuste do seu modelo traz mais informação do que o modelo somente com o intercepto.

Se o p-valor for menor do que o nível de significância estabelecido (alfa), rejeitamos **hipótese nula** e concluímos que o nosso modelo fornece um ajuste melhor do que o modelo somente com o intercepto. O modelo é significativo.

Teste F

OLS Regression Results

Dep. Variable:	Fatura	**R-squared:**	0.747
Model:	OLS	**Adj. R-squared:**	0.745
Method:	Least Squares	**F-statistic:**	301.7
Date:	Sun, 03 Jul 2022	**Prob (F-statistic):**	4.74e-91
Time:	16:08:15	**Log-Likelihood:**	-2094.1
No. Observations:	310	**AIC:**	4196.
Df Residuals:	306	**BIC:**	4211.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-160.8979	48.390	-3.325	0.001	-256.117	-65.679
Estudante[T.Yes]	394.7980	36.106	10.934	0.000	323.750	465.846
Limite	0.1728	0.006	29.182	0.000	0.161	0.184
Idade	-2.9817	0.695	-4.288	0.000	-4.350	-1.614

Omnibus:	6.223	**Durbin-Watson:**	1.978
Prob(Omnibus):	0.045	**Jarque-Bera (JB):**	6.393
Skew:	-0.340	**Prob(JB):**	0.0409
Kurtosis:	2.818	**Cond. No.**	2.43e+04

P-valor menor do que 0.05, rejeitamos **hipótese nula** e concluímos que o nosso modelo fornece um ajuste melhor do que o modelo somente com o intercepto.

Suposições (restrições) para usar o modelo de Regressão linear

- Erros são independentes. **(Independência)**
- Erros seguindo uma distribuição normal **(Normalidade)**
- Erros com variância constante. **(Homocedasticidade)**
- Relação linear entre Y e os X. **(Linearidade)**
- Ausência de **multicolinearidade** (uma variável x explicativa não relacionada com a outra).

Independência dos erros

O modelo supõe que os erros são independentes entre si, um erro não tem correlação com o outro.

Como avaliar?

- **<u>Teste Durbin-Watson:</u>** estatística de correlação serial dos resíduos (varia de 0 a 4).

  - ❖ Próximo de 0 → correlação (+)
  - ❖ Próximo de 2 → correlação nula (ideal)
  - ❖ Próximo de 4 → correlação (-)

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

- Valores entre 1.5 e 2.5 são considerados não correlacionados.

Independência

```
# MODELO DE REGRESSÃO LINEAR MULTIPLA

modelo = smf.ols(formula = 'Fatura ~  Limite +  Cartoes + Renda + Estudante  ', data = credit)
modelo_fit = modelo.fit()
modelo_fit.summary()
```

OLS Regression Results

Dep. Variable:	Fatura	**R-squared:**	0.998
Model:	OLS	**Adj. R-squared:**	0.998
Method:	Least Squares	**F-statistic:**	3.293e+04
Date:	Sun, 03 Jul 2022	**Prob (F-statistic):**	0.00
Time:	19:12:49	**Log-Likelihood:**	-1366.4
No. Observations:	310	**AIC:**	2743.
Df Residuals:	305	**BIC:**	2762.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-754.0254	4.609	-163.590	0.000	-763.095	-744.955
Estudante[T.Yes]	502.1046	3.505	143.246	0.000	495.207	509.002
Limite	0.3272	0.001	318.664	0.000	0.325	0.329
Cartoes	24.4434	0.800	30.573	0.000	22.870	26.017
Renda	-10.1344	0.055	-183.544	0.000	-10.243	-10.026

Omnibus:	3.184	**Durbin-Watson:**	1.963
Prob(Omnibus):	0.203	**Jarque-Bera (JB):**	2.358
Skew:	-0.030	**Prob(JB):**	0.308
Kurtosis:	2.577	**Cond. No.**	2.47e+04

- Valor próximo de 2. A suposição de independência dos resíduos não foi violada.

Normalidade

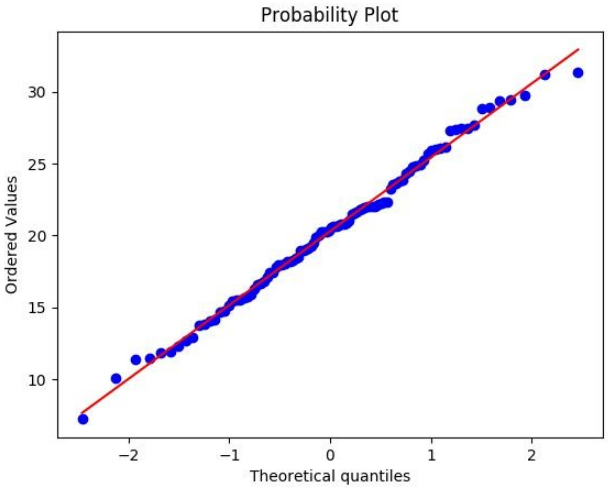
$$\begin{cases} H_0 = \text{O dado possui dist. Normal} \\ H_1 = \text{O dado não possui dist. Normal} \end{cases}$$

Testes:

- Omnibus test
- Jarque-Bera
- Kolmogorov-Smirnov
- Anderson-Darling
- Shapiro-Wilk

Se o <u>p-valor for maior do que o nível de significância</u> estabelecido (alfa), <u>não rejeitamos H_0,</u> os resíduos possuem dist. normal.

qqplot:



<u>**Modelo bem ajustado:**</u> pontos alinhados na reta vermelha.

Normalidade

```
# MODELO DE REGRESSÃO LINEAR MULTIPLA

modelo = smf.ols(formula = 'Fatura ~  Limite +  Cartoes + Renda + Estudante  ', data = credit)
modelo_fit = modelo.fit()
modelo_fit.summary()
```

OLS Regression Results

Dep. Variable:	Fatura	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.293e+04
Date:	Sun, 03 Jul 2022	Prob (F-statistic):	0.00
Time:	19:12:49	Log-Likelihood:	-1366.4
No. Observations:	310	AIC:	2743.
Df Residuals:	305	BIC:	2762.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-754.0254	4.609	-163.590	0.000	-763.095	-744.955
Estudante[T.Yes]	502.1046	3.505	143.246	0.000	495.207	509.002
Limite	0.3272	0.001	318.664	0.000	0.325	0.329
Cartoes	24.4434	0.800	30.573	0.000	22.870	26.017
Renda	-10.1344	0.055	-183.544	0.000	-10.243	-10.026

Omnibus:	3.184	Durbin-Watson:	1.963
Prob(Omnibus):	0.203	Jarque-Bera (JB):	2.358
Skew:	-0.030	Prob(JB):	0.308
Kurtosis:	2.577	Cond. No.	2.47e+04

- P-valor maior do que 0.05, não rejeitamos a hipótese nula. Os resíduos possuem distribuição normal.

Suposições do modelo

Linearidade, homocedasticidade, multicolinearidade e pontos discrepantes

Suposição de Linearidade

- Uma das suposições do modelo linear, é que existe uma relação linear entre x e y.

- Y está linearmente relacionado a X se a taxa de variação de Y em relação a X for constante (independente do valor de X).

Isso significa que:

A linearidade está nos parâmetros $\beta' s$.

- **Não podemos ter modelos do tipo:**

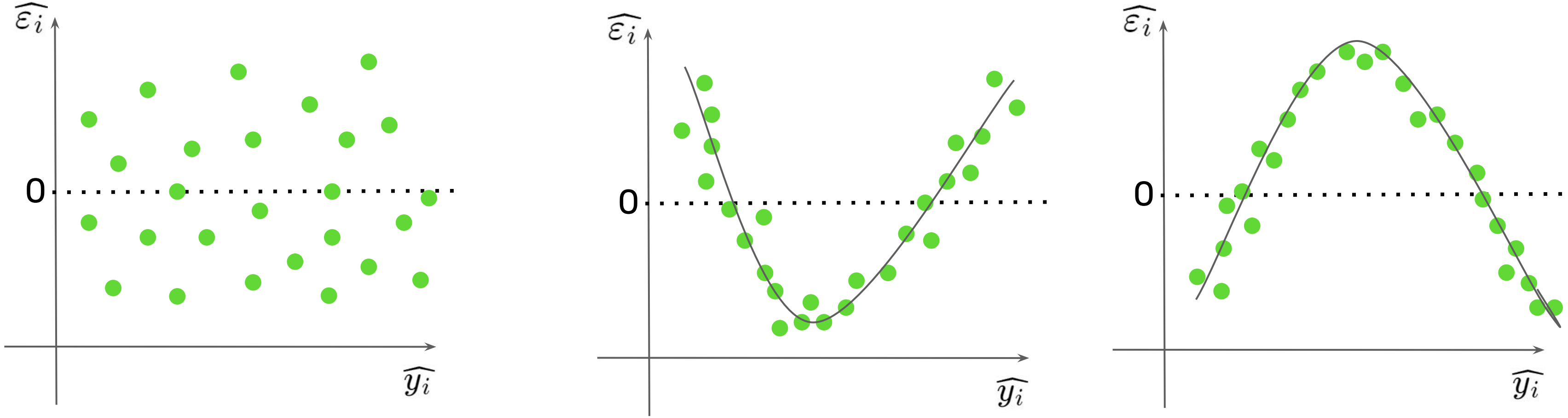
$$Y = \beta_0 + \beta_1^2 x_1 \qquad\qquad Y = \beta_0 + log(\beta_1)x_1$$

- **Podemos ter modelos do tipo:**

$$Y = \beta_0 + \beta_1 x_1^2 \qquad Y = \beta_0 + \beta_1 x_1^3 \qquad Y = \beta_0 + \beta_1 log(x_1)$$

Avaliando a Suposição de Linearidade

- **Gráfico Resíduos vs. Valores ajustados.**



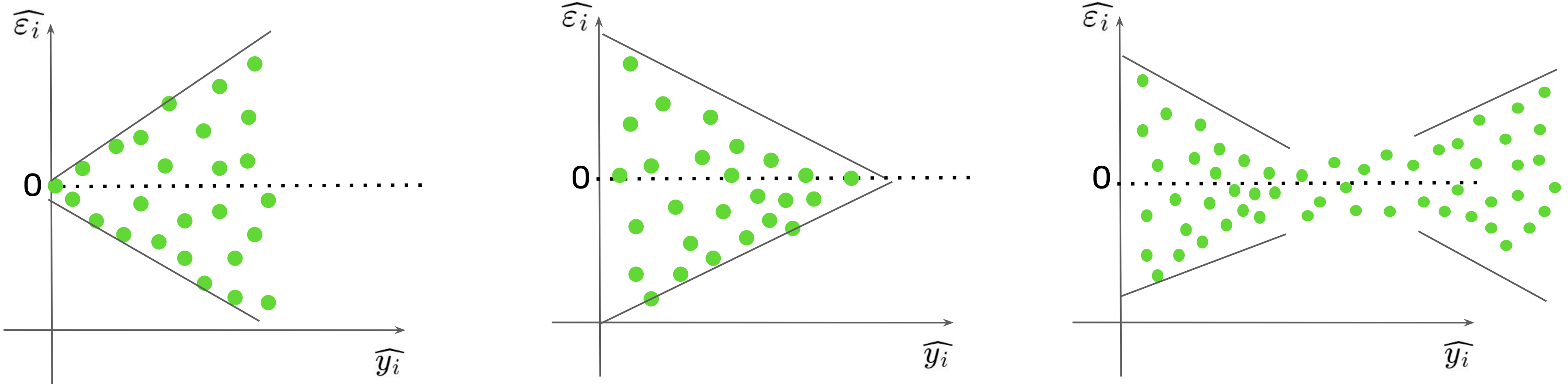
<u>Modelo bem ajustado:</u> Resíduos dispersos aleatoriamente em torno de zero (pontos distribuídos sem um padrão aparente).

<u>Problemas de linearidade:</u> As curvaturas indicam que a suposição de linearidade está sendo violada.

Homocedasticidade (Variância Constante)

- Supõe que a variância do erro cometido pelo modelo (valor real - valor estimado) é constante em todos os valores do Y.

Gráfico Resíduos vs. Valores ajustados.



<u>**Problemas de Heterocedasticidade:**</u> Dispersão dos resíduos aumenta ou diminui conforme o valor do predito, comum quando a variável resposta refere-se a contagens.

Homocedasticidade (Variância Constante)

$$\begin{cases} H_0 \text{ : A variância dos erros é constante (Homocedásticos)} \\ H_1 \text{ : A variância dos erros não é constante (Heterocedásticos)} \end{cases}$$

- **Testes**
  - White
  - BP (Breusch-Pagan)
  - Goldfeld-Quandt

Se o p-valor for maior do que o nível de significância estabelecido (alfa), não rejeitamos H_0, os resíduos possuem variância constante.

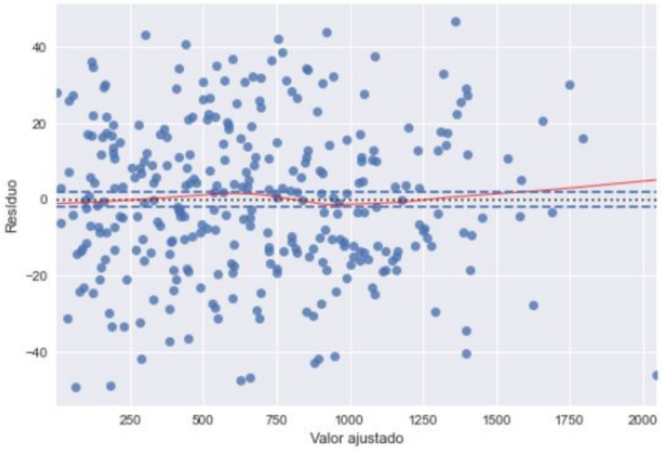
```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
from statsmodels.stats.diagnostic import het_white

#Heteroskedasticity tests

#Breush-Pagan test:

name = ["Breusch pagan statistic", "p-value"]
test = sms.het_breuschpagan(modelo_residuo, modelo3_fit.model.exog)
print(lzip(name, test))
print()

#Goldfeld-Quandt test

name = ["Goldfeld-Quandt - statistic", "p-value"]
test = sms.het_goldfeldquandt(modelo_residuo, modelo3_fit.model.exog)
print(lzip(name, test))
print()

#White's Test
name = ["White's statistic", "p-value"]
test = het_white(modelo_residuo, modelo3_fit.model.exog)
print(lzip(name, test))
print()
```

```
[('Breusch pagan statistic', 11.440451366646476), ('p-value', 0.022035423850913942)]

[('Goldfeld-Quandt - statistic', 1.0685287103201158), ('p-value', 0.3426711962435063)]

[('White's statistic', 15.88401260701602), ('p-value', 0.2554499027961704)]
```

Linearidade e Homocedasticidade

Gráfico Resíduos vs. Valores ajustados

```
# GRÁFICO PARA LINEARIDADE E HOMOCEDASTICIDADE #

modelo_residuo = modelo_fit.resid
modelo_y_ajustado = modelo_fit.fittedvalues

sns.residplot(modelo_y_ajustado, modelo_residuo,
                          lowess=True,
                          line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8})
plt.axhline(y = 2, linestyle='dashed')
plt.axhline(y = -2, linestyle='dashed')
plt.ylabel('Resíduo')
plt.xlabel('Valor ajustado')
plt.show()
```



Ausência de Multicolinearidade

Multicolinearidade indica que as variáveis explicativas são altamente correlacionadas.

Detecção:

- Calcular a matriz de correlação das variáveis explicativas e verificar se existem coeficientes de correlação elevados.

- Calcular o VIF(Variance Inflation Factors) : Quanto maior o VIF, mais grave é a multicolinearidade.

```
# MODELO DE REGRESSÃO LINEAR MULTIPLA

modelo = smf.ols(formula = 'Fatura ~  Limite +  Cartoes + Renda + Estudante  ', data = credit)
modelo_fit = modelo.fit()
modelo_fit.summary()
```

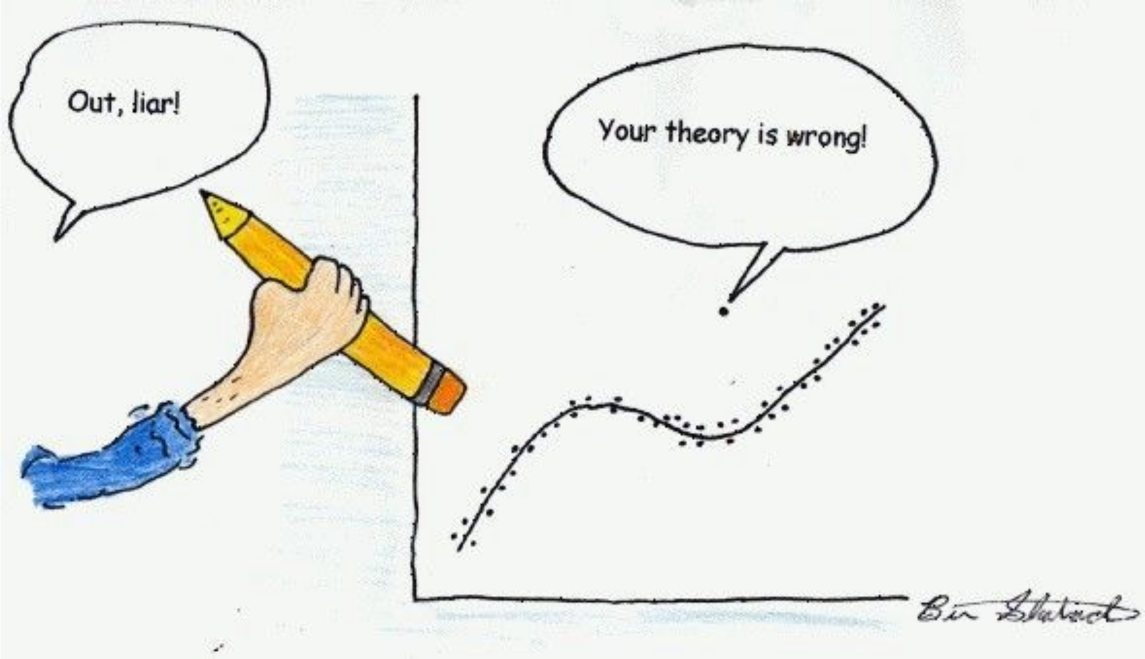
VIF	As variáveis explicativas são...
1<VIF<5	Não correlacionadas / fracamente correlacionadas
5<VIF<10	Moderadamente correlacionadas
VIF >10	Fortemente correlacionadas

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = pd.DataFrame()

vif["VIF Factor"] = [variance_inflation_factor(modelo_fit.model.exog, i) for i in range(1, modelo_fit.model.exog.shape[1])]
vif["Variable"]   = modelo_fit.model.exog_names[1:]
print(vif)
```

```
   VIF Factor          Variable
0    1.044546  Estudante[T.Yes]
1    3.423107            Limite
2    1.002592           Cartoes
3    3.371060             Renda
```

Outliers

- São aquelas observações que se destacam de todo o conjunto de observações, por terem seu valor muito afastado dos demais, fornecendo um valor residual muito fora dos padrões apresentados pela maioria das observações.

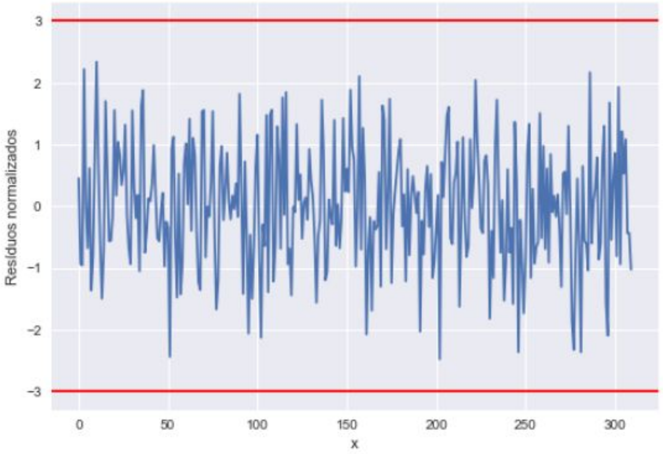


Detecção de outliers

Modelo bem ajustado:

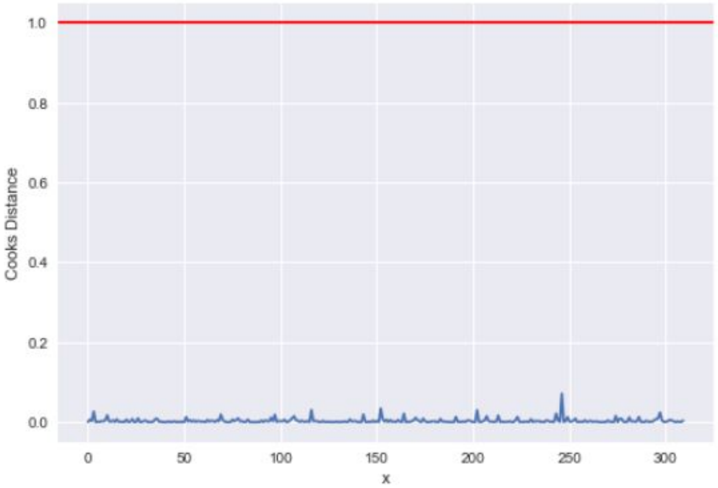
- Resíduos estudentizados entre -3 e 3.

```
from statsmodels.graphics.regressionplots import *

influence = modelo_fit.get_influence()
inf_sum = influence.summary_frame().round(3)

Student_resid = influence.resid_studentized
plt.plot(Student_resid)
plt.xlabel('x')
plt.axhline(y=-3, color='r', linestyle='-')
plt.axhline(y=3, color='r', linestyle='-')
plt.ylabel('Resíduos normalizados')
plt.show()
```



- Distância de Cook menor que 1

```
cooks = influence.cooks_distance
X = credit[['Cartoes', 'Estudante','Renda', 'Limite']]

plt.plot(cooks[0])
plt.xlabel('x')
plt.ylabel('Cooks Distance')
plt.axhline(y=1, color='r', linestyle='-')
plt.show()
```



Alternativas para a violação dos pressupostos da regressão linear

Análise de resíduos

- Os resíduos contém informação sobre o motivo do modelo não ter se ajustado bem aos dados. Eles conseguem indicar se uma ou mais suposições do modelo foram violadas.

- Principais problemas detectados através da análise dos resíduos:
  - Não-linearidade da relação entre X e Y ;
  - Não normalidade dos erros;
  - Variância não-constante dos erros (heterocedasticidade);
  - Correlação entre os erros;
  - Presença de outliers ou observações atípicas.

Violação da Normalidade

Alternativas:

- Utilizar outras classes de modelos (ex: Modelos Lineares Generalizados), que tem outros tipos de distribuição, Poisson, Gamma, Binomial, modelos zero inflado e etc.

- Transformação da variável resposta, como a transformação de Box-Cox:

$$y_t = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, se\ \lambda \neq 0 \\ \log(y), \quad se\ \lambda = 0 \end{cases}$$

- Consiste em estimar o parâmetro λ da transformação Y .
- Os valores mais comuns de λ correspondem às seguintes transformações:

$\hat{\lambda}$	**Transformação**
-1,0	$Y' = Y^{-1} = 1/Y$
-0,5	$Y' = Y^{-0,5} = 1/\sqrt{Y}$
0,5	$Y' = Y^{0,5} = \sqrt{Y}$
1,0	$Y' = Y^{1} = Y$

- O método de estimação de λ é bastante trabalhoso e, por isso, é realizado computacionalmente.

Violação do pressuposto de independência

Exemplos:

- **Medidas repetidas** ⇒ coleta-se a medida em um mesmo indivíduo em diferentes instantes de tempo.

- **Série temporal** ⇒ os dados possuem estrutura temporal que não é captada pelo modelo.

- **Dados hierárquicos** ⇒ indivíduos agrupados, por exemplo, alunos em uma escola.

Em caso de observações dependentes, optar por técnicas e modelos que incorporem a correlação entre as observações.

- ○ Medidas ao longo do tempo : Modelos de séries temporais (AR, ARMA, VAR e etc).

- ○ Medidas repetidas no mesmo indivíduo: modelos longitudinais, hierárquicos, análise de sobrevivência.

Heterocedasticidade (variância não constante)

Alternativas:

- Transformações da variável resposta (Y).
- Optar por outro método de estimação (Mínimos quadrados ponderados (WLS), estimadores de erros padrão robusto).
- Modelos hierárquicos.
- Técnicas de reamostragem para estimar os erros.

Heterocedasticidade (variância não constante)

- A suposição de normalidade e variância não constante andam de mãos dadas. Portanto as transformações para estabilizar a variância são aplicadas na variável resposta (Y) e podem também ajudar no pressuposto de normalidade dos resíduos.

Entre as transformações possíveis, temos:

- **Raiz quadrada**

$$y' = \sqrt{y}$$

- **Inverso**

$$y' = \frac{1}{y}$$

- **Logaritmo natural (base e)**

$$y' = ln(y)$$

Estabiliza a variância quando ela tende a crescer à medida que y cresce. Pode ajudar a normalizar os dados.

- **Quadrática**

$$y' = y^2$$

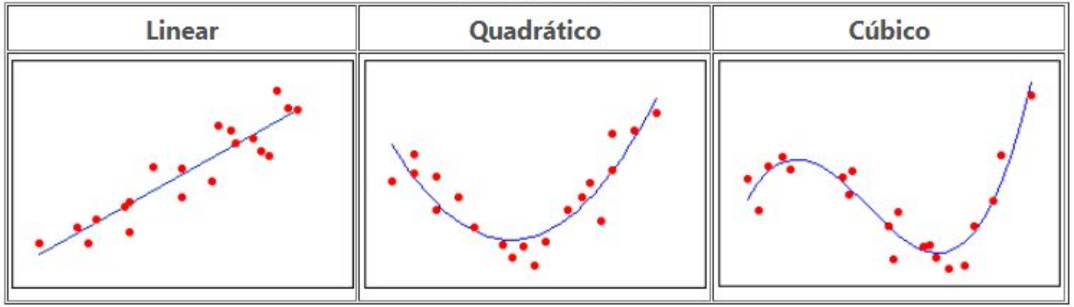
Normaliza os dados quando os resíduos possuem assimetria negativa.

Violação da Linearidade

Alternativas:

- Ajustar modelo com polinômios (Ex: incluir X^2).

A maneira mais comum de ajustar curvas aos dados usando regressão linear é incluir termos polinomiais, como variáveis explicativas quadradas ou cúbicas .



- Transformação na variável explicativa (X).
- Optar por outra classe de modelos (GAMs, Boosting de modelos lineares).
- Splines (é uma função definida por partes por polinômios).

Violação da Linearidade

- Para corrigir o problema de não linearidade as transformações são aplicadas na variável explicativa (X).

Entre as transformações possíveis, temos:

- **Raiz quadrada**

$$x' = \sqrt{x}$$

- **Inverso**

$$x' = \frac{1}{x}$$

- **Exponencial**

$$x' = exp(x)$$

- **Logaritmo natural (base e)**

$$x' = ln(x)$$

- **Quadrática**

$$x' = x^2$$

Violação da ausência de multicolinearidade

Alternativas:

- Eliminar do modelo, uma por vez, as variáveis que estão correlacionadas com outras já existentes.

- Reespecificar o modelo. Se X1, X2 e X3 são correlacionadas, pode-se tentar encontrar uma relação entre elas, com sentido prático, para tentar preservar a informação das variáveis originais. Ex: (X1+X2)/X3 ou podemos usar termos de interação(X1*X2*X3).

- Regressão corrigida (ridge regression) e componentes principais.

- Centralizar as variáveis no caso de regressão polinomial.

Violação da presença de pontos influentes

Alternativas:

- Descartar a observação se esta for resultado de algum erro de medição, digitação etc.

- Utilizar um método de estimação que não seja tão influenciado por tais observações quanto o método de mínimos quadrados (ex: métodos de regressão robusta).

Sobre as transformações

- Quando fazemos uma transformação em Y, as estimações e previsões estão expressas em novas unidades, conforme a transformação utilizada. Portanto, a interpretação do modelo deve ser feita na escala transformada.

- Adicionar transformações aumenta a complexidade do modelo e a sua interpretação.

- A escolha da melhor transformação é feita, de certa forma, empiricamente, pois **não há garantias de que certa transformação solucionará o problema detectado.**

- A transformação deve ser feita, respeitando a característica do dado. Por exemplo, não podemos aplicar transformação quadrática em dados negativos.

Sobre as transformações

- Transformar a variável resposta afeta a variância, normalidade e linearidade.
- Transformar a variável explicativa só afeta linearidade.
- Então devemos primeiro trabalhar com a variância:
  - Transformar a variável Y e corrigir a variância e normalidade.
  - Em seguida corrigir a linearidade: transformar a variável explicativa.

RESUMO E TAKEAWAYS

TAKEAWAY **#1**

Os modelos de regressão linear são considerados adequados e bem ajustados quando todas as suposições do modelo são atendidas.

TAKEAWAY **#2**

As suposições do modelo são:

- Observações são independentes. (Independência)
- Erros seguindo uma distribuição normal (Normalidade)
- Erros com variância constante. (Homocedasticidade)
- Relação linear entre Y e os X. (Linearidade)
- Ausência de multicolinearidade (uma variável x explicativa não relacionada com a outra).

TAKEAWAY **#3**

O modelo de Regressão linear é um modelo que retorna uma média.

Qual será em média o valor da fatura, para alguém que possui um limite de $1000?

Precisamos nos preocupar com valores discrepantes nos dados, valores discrepantes influenciam a média.

T