# Ciência dos dados

## Modelo de regressão `statsmodels.OLS`

Insper

Magalhães e Lima (7ª. Edição): Seção 9.5

Montgomery. Estatística Aplicada e Probabilidade para Engenheiros: Capítulo 11

# Entendendo a saída do statsmodels.OLS

# Problema - GAPMinder



**Problema**: Considere que o objetivo aqui seja explicar/prever a emissão de gás carbono (CO2) per capita de um país em função da renda (PIB) per capita.

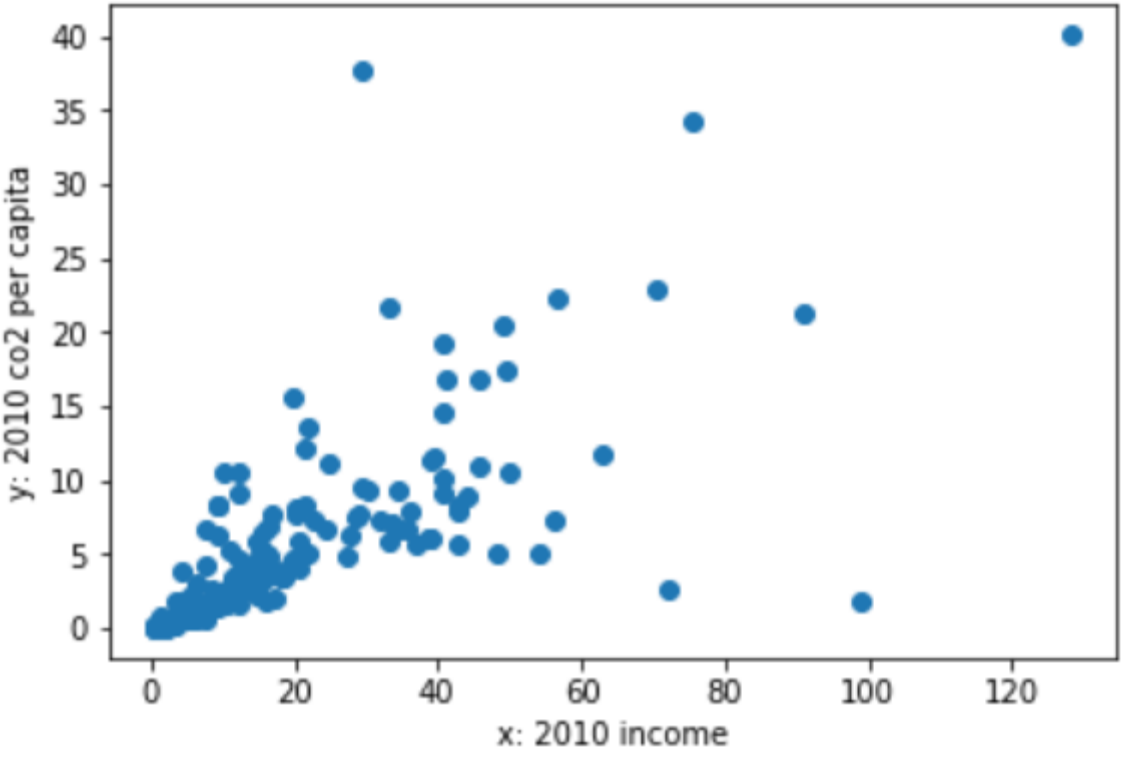
$y$**: variável dependente (resposta)** – emissão de CO2.

$x$**: variável independente (explicativa) –** renda (PIB) per capita.

**Modelo de regressão linear simples**: modelo que associa $y$ em função de uma variável explicativa $x$.

# Problema - GAPMinder





# Problema - GAPMinder



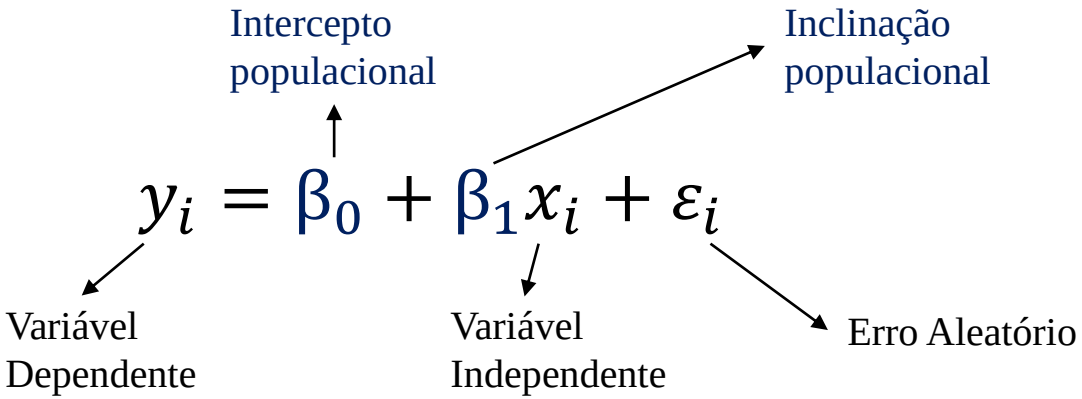
Modelo geral:

$$\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$$

Problema:

$$\mathrm{E}(\text{emissão CO2}|Renda) = \beta_0 + \beta_1 Renda$$

**Significado dos termos do modelo geral:**



# Teste de hipóteses



Usualmente, uma das hipóteses em análise de regressão é avaliar a significância da regressão.

$$\begin{matrix} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{matrix} \rightarrow \begin{matrix} H_0: \text{não há relação entre } x \text{ e } Y \\ H_1: \text{há relação entre } x \text{ e } Y \end{matrix}$$

Para realizar esse teste de hipóteses, será necessário atribuir distribuição aos erros $\varepsilon_i$, além de outras suposições ao modelo.

# Suposições do modelo

**Como verificar essas suposições?**

- Os **erros têm distribuição normal** com média e variância constante, ou seja, $\varepsilon_i \sim N(0,\sigma^2)$.

**Teste Omnibus e Teste Jarque-Bera**

- Os **erros são independentes** entre si, ou seja, $\mathrm{Corr}(\varepsilon_i, \varepsilon_j)=0$, para qualquer $i \neq j$.

**Teste Durbin-Watsom**

- **O modelo é linear nos parâmetros**.

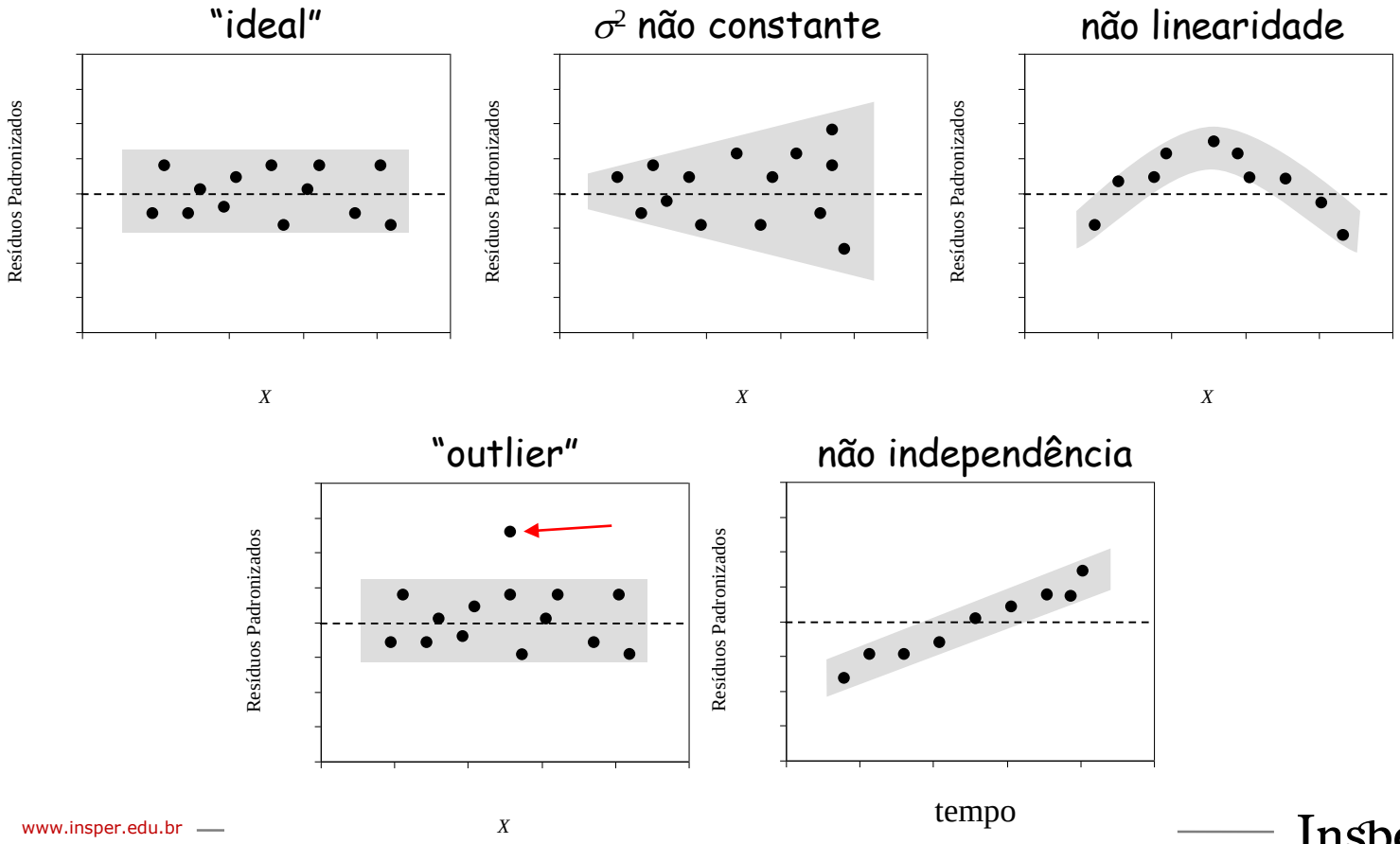
**Montgomery e Runger (2018, Seção 11-7.1 e Seção 11.7-2)**

- **Homocedasticidade**: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ para qualquer $i = 1, \ldots, n$.

**Montgomery e Runger (2018, Seção 11-7.1)**

# Análise de resíduos





# statsmodels.OLS

Considerando exemplo do GAPMinder

$x$: Renda (PIB) per capita

$y$: emissão de CO2

# Biblioteca `statsmodels`



**Statsmodels** para modelagem estatística.

"É um módulo Python que permite aos usuários explorar os dados, estimar modelos estatísticos, e realizar testes estatísticos. Uma extensa lista de estatística descritiva, testes estatísticos, funções de plotagem e estatísticas de resultados estão disponíveis para diferentes tipos de dados."

Fonte: https://www.vooo.pro/insights/um-tutorial-completo-para-aprender-data-science-com-python-do-zero/

# `Statsmodels.OLS`



A seguir, vamos compreender os resultados (`summary`) obtidos do ajuste de regressão utilizando o Método dos Mínimos Quadrados (MMQ). Também chamado de Mínimos Quadrados Ordinários (MQO) ou OLS (do inglês Ordinary Least Squares).

# Statsmodels.OLS



```
import statsmodels.api as sm # Importe da biblioteca

x = df['2010_income'] # Definindo renda como explicativa
y = df['2010_co2']    # Definindo CO2 como resposta

xc = sm.add_constant(x) # Adiciona coluna de 1s para estimar intercepto
model = sm.OLS(y,xc)    # Define o modelo
results = model.fit()   # Faz o ajuste
results.summary()       # Mostra os resultados
```

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | 2010_co2 | **R-squared:** | 0.561 |
| **Model:** | OLS | **Adj. R-squared:** | 0.559 |
| **Method:** | Least Squares | **F-statistic:** | 235.2 |
| **Date:** | Wed, 30 Oct 2019 | **Prob (F-statistic):** | 9.80e-35 |
| **Time:** | 23:16:45 | **Log-Likelihood:** | -534.72 |
| **No. Observations:** | 186 | **AIC:** | 1073. |
| **Df Residuals:** | 184 | **BIC:** | 1080. |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.6557 | 0.423 | 1.551 | 0.123 | -0.178 | 1.490 |
| **2010_income** | 0.2433 | 0.016 | 15.337 | 0.000 | 0.212 | 0.275 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 100.299 | **Durbin-Watson:** | 2.049 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2384.399 |
| **Skew:** | 1.435 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 20.304 | **Cond. No.** | 35.7 |

# Statsmodels.OLS

- Nesses resultados, deve-se entender se Renda (PIB) de um país é ou não relevante para explicar emissão de CO2.

- Ainda, é necessário validar o modelo avaliando se suas suposições são válidas.

# Statsmodels.OLS

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | 2010_co2 | R-squared: | 0.561 |
| Model: | OLS | Adj. R-squared: | 0.559 |
| Method: | Least Squares | F-statistic: | 235.2 |
| Date: | Wed, 30 Oct 2019 | Prob (F-statistic): | 9.80e-35 |
| Time: | 23:16:45 | Log-Likelihood: | -534.72 |
| No. Observations: | 186 | AIC: | 1073. |
| Df Residuals: | 184 | BIC: | 1080. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.6557 | 0.423 | 1.551 | 0.123 | -0.178 | 1.490 |
| 2010_income | 0.2433 | 0.016 | 15.337 | 0.000 | 0.212 | 0.275 |

| | | | |
|---|---|---|---|
| Omnibus: | 100.299 | Durbin-Watson: | 2.049 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 2384.399 |
| Skew: | 1.435 | Prob(JB): | 0.00 |
| Kurtosis: | 20.304 | Cond. No. | 35.7 |

## Teste t : valor *p*



$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

Coluna mostra as estimativas dos coeficientes da reta:

**const** ➔ estimativa do intercepto $\hat{\beta}_0$

**2010_income** ➔ estimativa da inclinação da reta $\hat{\beta}_1$

# Statsmodels.OLS



OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | 2010_co2 | **R-squared:** | 0.561 |
| **Model:** | OLS | **Adj. R-squared:** | 0.559 |
| **Method:** | Least Squares | **F-statistic:** | 235.2 |
| **Date:** | Wed, 30 Oct 2019 | **Prob (F-statistic):** | 9.80e-35 |
| **Time:** | 23:16:45 | **Log-Likelihood:** | -534.72 |
| **No. Observations:** | 186 | **AIC:** | 1073. |
| **Df Residuals:** | 184 | **BIC:** | 1080. |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.6557 | 0.423 | 1.551 | 0.123 | -0.178 | 1.490 |
| **2010_income** | 0.2433 | 0.016 | 15.337 | 0.000 | 0.212 | 0.275 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 100.299 | **Durbin-Watson:** | 2.049 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2384.399 |
| **Skew:** | 1.435 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 20.304 | **Cond. No.** | 35.7 |

## Teste t : valor *p*

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

Coluna mostra o **valor-p** do Teste t que possui $H_0$: $\beta_i = 0$

**Regra geral➔**
se valor-p < $\alpha$, então rejeita-se a hipótese nula.

Se teste for para $\beta_1$, mostra que a variável explicativa (Renda) é relevante para explicar mudanças na variável resposta (CO2).

# Statsmodels.OLS

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | 2010_co2 | **R-squared:** | 0.561 |
| **Model:** | OLS | **Adj. R-squared:** | 0.559 |
| **Method:** | Least Squares | **F-statistic:** | 235.2 |
| **Date:** | Wed, 30 Oct 2019 | **Prob (F-statistic):** | 9.80e-35 |
| **Time:** | 23:16:45 | **Log-Likelihood:** | -534.72 |
| **No. Observations:** | 186 | **AIC:** | 1073. |
| **Df Residuals:** | 184 | **BIC:** | 1080. |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.6557 | 0.423 | 1.551 | 0.123 | -0.178 | 1.490 |
| **2010_income** | 0.2433 | 0.016 | 15.337 | 0.000 | 0.212 | 0.275 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 100.299 | **Durbin-Watson:** | 2.049 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2384.399 |
| **Skew:** | 1.435 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 20.304 | **Cond. No.** | 35.7 |

## Teste Omnibus

Teste a normalidade dos resíduos:

$H_0$ : a distribuição dos resíduos é normal

$H_1$ : a distribuição dos resíduos não é normal

**Prob(Omnibus) ➔**
é o valor-p desse teste de normalidade

**Regra geral➔**

- (IDEAL) **Prob(Omnibus)** > $\alpha$
- Se **Prob(Omnibus)** for muito baixo (menor do que < $\alpha$), então existe evidência de que os resíduos **não são** distribuídos normalmente, violando nesse caso essa suposição do modelo de regressão.

# Statsmodels.OLS

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | 2010_co2 | **R-squared:** | 0.561 |
| **Model:** | OLS | **Adj. R-squared:** | 0.559 |
| **Method:** | Least Squares | **F-statistic:** | 235.2 |
| **Date:** | Wed, 30 Oct 2019 | **Prob (F-statistic):** | 9.80e-35 |
| **Time:** | 23:16:45 | **Log-Likelihood:** | -534.72 |
| **No. Observations:** | 186 | **AIC:** | 1073. |
| **Df Residuals:** | 184 | **BIC:** | 1080. |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.6557 | 0.423 | 1.551 | 0.123 | -0.178 | 1.490 |
| **2010_income** | 0.2433 | 0.016 | 15.337 | 0.000 | 0.212 | 0.275 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 100.299 | **Durbin-Watson:** | 2.049 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2384.399 |
| **Skew:** | 1.435 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 20.304 | **Cond. No.** | 35.7 |

## Teste Jarque-Bera

Outro teste de normalidade dos resíduos:

$H_0$ : a distribuição dos resíduos é normal

$H_1$ : a distribuição dos resíduos não é normal

**Prob(JB) ➔**
é o valor-p desse teste de normalidade

**Regra geral➔**
- (IDEAL) **Prob(JB)** > $\alpha$
- Se **Prob(JB)** for muito baixo (menor do que < $\alpha$), então existe evidência de que os resíduos **não são** distribuídos normalmente, violando nesse caso essa suposição do modelo de regressão.

Statsmodels.OLS

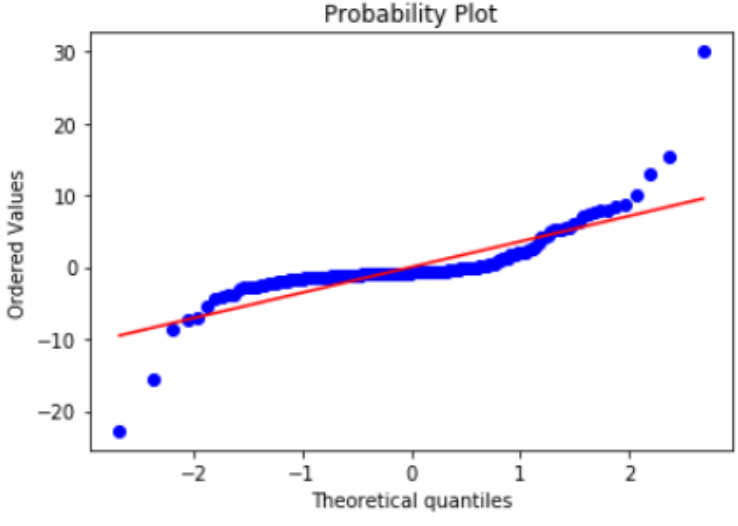
RESÍDUOS

## Resíduos

Plot da normalidade dos resíduos - recurso descritivo.

Esse plot pode ser feito usando o atributo `resid` dos resultados da regressão.

```
stats.probplot(results.resid, dist="norm", plot=plt);
```



# Statsmodels.OLS

## HOMOCEDASTICIDADE

**Análise de homocedasticidade**
Verifique visualmente se a hipótese de homocedasticidade é válida.

Resp:
Para validar, descritivamente, indicativos na suposição de homocedasticidade considerando os resíduos obtidos do modelo ajustado, pode ser feito gráficos dos resíduos $\varepsilon_i = y_i - \hat{y}_i = y_i$ contra os valores preditos $\hat{y}_i$ e contra cada variável independente $x_i$ , com $\$i = 1, \ldots, n$. Vide Montgomery e Runger (2018), Seção 11-7.1, página 351, para mais detalhes.

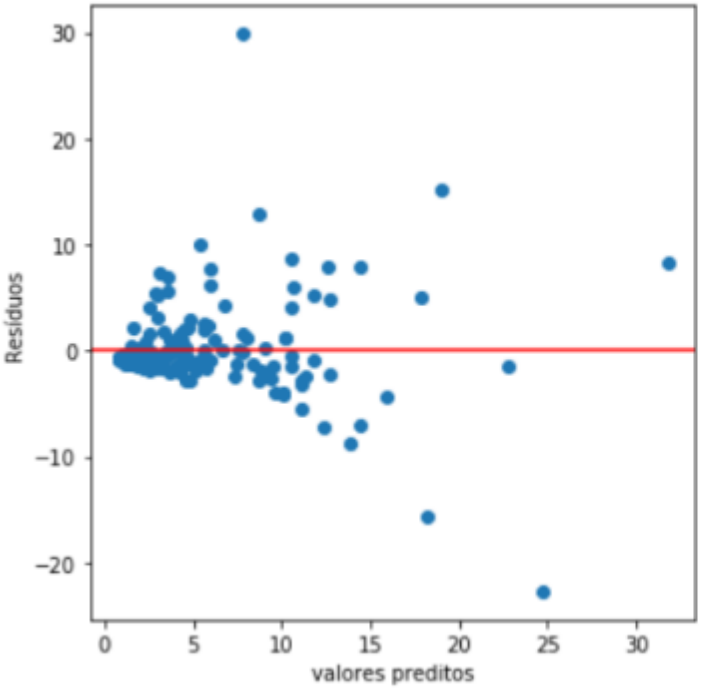
Interprentando os gráficos abaixo, nota-se que a variância dos resíduos está crescendo com a magnitiude de $y_i$ e com a magnitude de $y_i$ dando indicativo que a suposição de homocedasticidade deva estar violada. Uma solução pode ser a transformação nas variáveis dependente e/ou independentes ou usar modelos de regressão mais robustos.

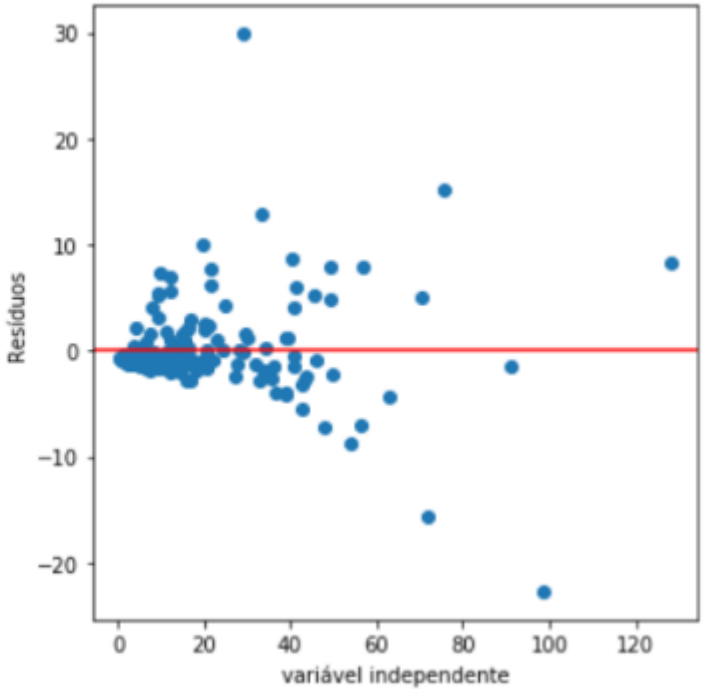
# Statsmodels.OLS



## Análise de homocedasticidade

Interprentando os gráficos abaixo, nota-se que a variância dos resíduos está crescendo com a magnitiude de $y_i$ e com a magnitude de $x_i$ dando indicativo que a suposição de homocedasticidade deva estar violada. Uma solução pode ser a transformação nas variáveis dependente e/ou independentes ou usar modelos de regressão mais robustos.



# Statsmodels.OLS



## Conclusão

A conclusão do ajuste anterior é que as suposições de normalidade (via teste de hipóteses) e de homocedasticidade (graficamente) não estão válidas.

Logo, o modelo ajustado que relacionada como emissão de CO2 de países pode ser explicado por Renda não deve ser considerado para tomada de decisões.

Uma solução pode ser a **transformação nas variáveis dependente e/ou independentes** ou usar modelos de regressão mais robustos.

# statsmodels.OLS

Considerando exemplo do GAPMinder

$\log x$: log(Renda (PIB) per capita)

$\log y$: log(emissão de CO2)

***Atenção:*** *para interpretação dos coeficientes, neste caso, consulte Montgomery & Runger (2018)*

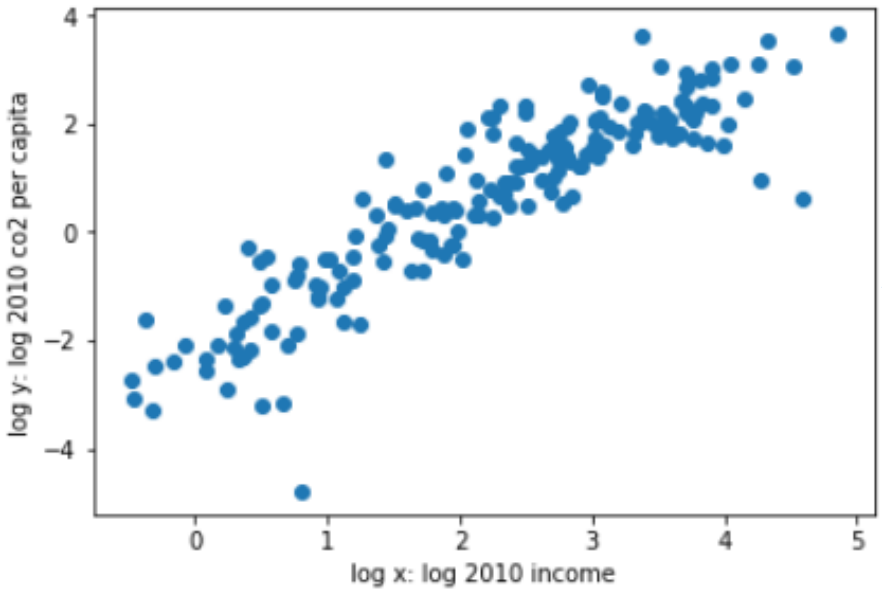
# Ajuste com uso da escala log nas variáveis



```
df['log_2010_income'] = np.log(df['2010_income'])
df['log_2010_co2'] = np.log(df['2010_co2'])

logx = df['log_2010_income'] # Definindo log(renda) como explicativa
logy = df['log_2010_co2']    # Definindo log(CO2) como resposta
```

```
plt.scatter(logx,logy);
plt.xlabel("log x: log 2010 income");
plt.ylabel("log y: log 2010 co2 per capita");
```




Ajuste com escala logaritmo natural nas variáveis:
Renda e emissão de CO2.

# Problema - GAPMinder



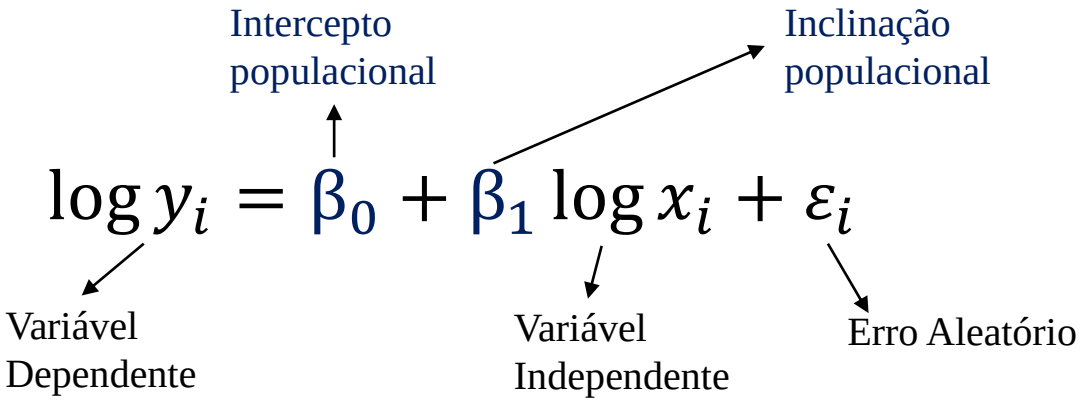
Modelo geral:

$$\mathrm{E}(\log Y \,|\log x) = \beta_0 + \beta_1 \log x$$

Problema:

$$\mathrm{E}(\log(\text{emissão CO2})|\,\log \text{Renda}) = \beta_0 + \beta_1 \log Renda$$

**Significado dos termos do modelo geral:**



# Ajuste com uso da escala log nas variáveis



```
# Ajuste considerando variáveis na escala log

logxc = sm.add_constant(logx) # Adiciona coluna de 1s para estimar intercepto
model = sm.OLS(logy,logxc)    # Define o modelo
results = model.fit()         # Faz o ajuste
results.summary()             # Mostra os resultados
```

Com a saída (próximo slide), avalie as suposições do modelo agora que as variáveis estão na escala log.

Verifique a significância da regressão.

Consulte aqui para estudar algumas transformações em variáveis.

# Ajuste com uso da escala log nas variáveis

- Nesses resultados, deve-se entender se log(Renda )de um país é ou não relevante para explicar o log(CO2).

- Ainda, é necessário validar o modelo avaliando se suas suposições são válidas.

- **Analise!**

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_2010_co2 | **R-squared:** | 0.819 |
| **Model:** | OLS | **Adj. R-squared:** | 0.818 |
| **Method:** | Least Squares | **F-statistic:** | 833.8 |
| **Date:** | Thu, 31 Oct 2019 | **Prob (F-statistic):** | 2.94e-70 |
| **Time:** | 00:43:43 | **Log-Likelihood:** | -200.52 |
| **No. Observations:** | 186 | **AIC:** | 405.0 |
| **Df Residuals:** | 184 | **BIC:** | 411.5 |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -2.0546 | 0.107 | -19.188 | 0.000 | -2.266 | -1.843 |
| **log_2010_income** | 1.2135 | 0.042 | 28.876 | 0.000 | 1.131 | 1.296 |

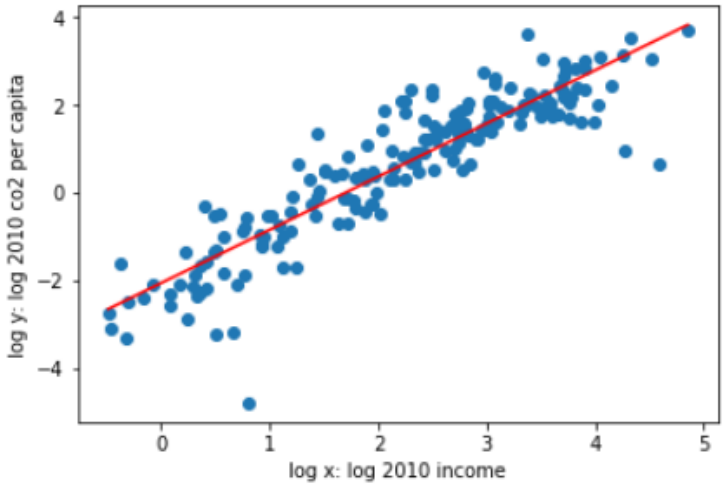
| | | | |
|---|---|---|---|
| **Omnibus:** | 52.051 | **Durbin-Watson:** | 2.134 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 209.893 |
| **Skew:** | -1.010 | **Prob(JB):** | 2.64e-46 |
| **Kurtosis:** | 7.796 | **Cond. No.** | 5.83 |

# Ajuste com uso da escala log nas variáveis

**Analise!**

```
logx_v = np.linspace(logx.min(), logx.max(), 500)
```

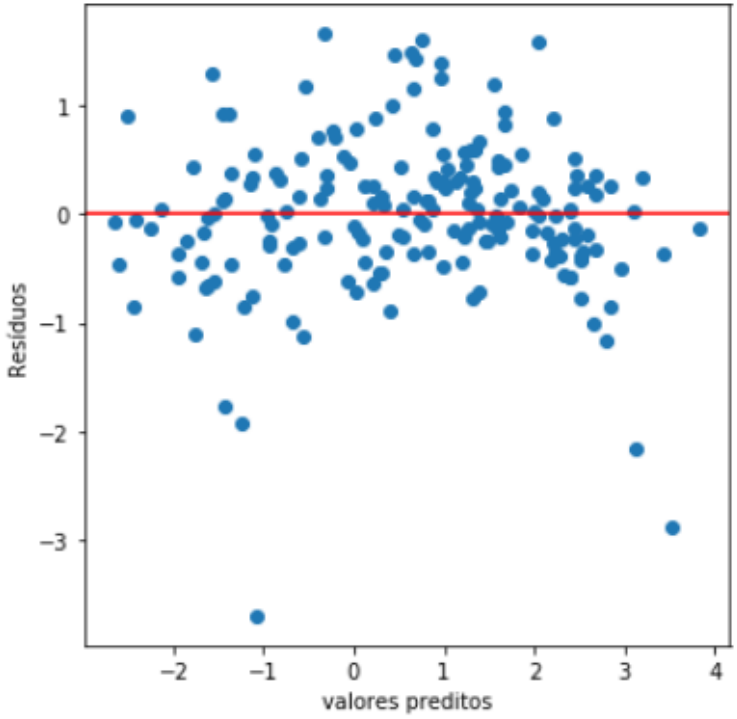
```
logx_vc = sm.add_constant(logx_v)
logy_vc = results.predict(logx_vc)
plt.scatter(logx,logy);
plt.plot(logx_v, logy_vc, color="r")
plt.xlabel("log x: log 2010 income");
plt.ylabel("log y: log 2010 co2 per capita");
```

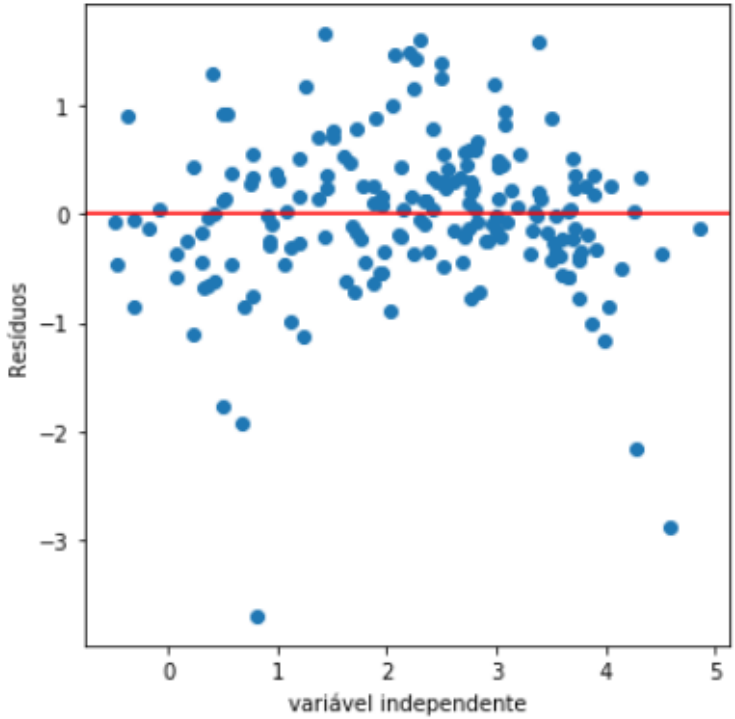


# Ajuste com uso da escala log nas variáveis



## Análise de homocedasticidade




Analise!

# Ajuste com uso da escala log nas variáveis



## Análise de homocedasticidade

```
fig = plt.figure(figsize=(10, 5))
plt.subplot(121)
plt.scatter(results.predict(logxc),results.resid); #logxc contem matriz de planejamento usada no ajuste OLS
plt.axhline(y=0, color='r', linestyle='-');
plt.ylabel('Resíduos')
plt.xlabel('valores preditos')

plt.subplot(122)
plt.scatter(logx,results.resid); #logx contem apenas a variável independente utilizada no ajuste linear
plt.axhline(y=0, color='r', linestyle='-');
plt.ylabel('Resíduos')
plt.xlabel('variável independente')

plt.tight_layout()
plt.show()
```

Insper

# Modelo de regressão MÚLTIPLA

# Um particular problema

arquivo ipynb

Atividade com contexto de regressão múltipla

# Atividade



- Download do notebook pelo Github
- Fazer individual e discutir em grupo:
- Usar arquivo:

`Aula27_Atividade_....ipynb`