Método OLS da Biblioteca statsmodels

Para termos acesso ao método OLS(Ordinary Last Squares), que é o Método dos Mínimos Quadrados, importaremos usando a seguinte linha de código.

```
import statsmodels.api as sm
```

Abaixo aplicaremos o método apenas para um dos produtos do nosso dataset, apenas para demonstrar o uso da biblioteca e interpretar os resultados da regressão linear.

Como chegar nos valores de x_values_laptop e y_values_laptop já foi discutido nas aulas anteriores, aqui filtramos para pegar esses valores apenas para o produto '12 MacBook (Mid 2017, Gold)' e renomeamos essa filtragem para xv_laptop e yv_laptop, usamos a função sm.add_constante(xv_laptop) para criarmos uma coluna de 1 que multiplicará com o resultado do coeficiente linear determinado pela regressão, precisa ser de 1 para não alterar o resultado. Esse método poderia ter sido adicionado de forma automática na função, mas a própria documentação indica usar esse método, portanto precisamos adicionar de forma manual.

```
xv_laptop = x_values_laptop['12 MacBook (Mid 2017, Gold)']
yv_laptop = y_values_laptop['12 MacBook (Mid 2017, Gold)']
X_laptop = sm.add_constant(xv_laptop)
model = sm.OLS(yv_laptop, X_laptop)
result = model.fit()
print(result.summary())
```

Finalmente definimos o modelo e usando o método fit() treinamos o método dos mínimos quadrados, esse método realiza os processos matemáticos discutidos na última aula determinando os coeficientes da equação que define a reta da equação linear. O método result.summary() nos apresenta os resultados obtidos para a regressão linear e vamos aprender a interpretá-los a seguir.

Interpretando .summary()

No início do sumário temos qual é a nossa **dependente variável** que é '12 MacBook (Mid 2017, Gold)' em seguida temos o **modelo** e **método** utilizado e ambos remetem ao método dos mínimos quadrados, após é apresentado a **data** e **horário** que foi criado o modelo. A seguir é apresentado o **número de observações** no dataset, **DF Residuals** é os graus de liberdade do nosso teste, determinado como o (número de observações – número de variáveis preditas – 1), **Df Model** é o número de variáveis preditas, **tipo de covariância** é a medida de como duas variáveis estão relacionadas, podendo ser positiva ou negativa, pode minimizar ou eliminar variáveis. **R^2** é o quanto a nossa variável independente é explicada pela nossa

variável dependente, expressa percentualmente, no nosso caso é 0,081, ou seja, 8,1% e o valor **ajustado de R^2** é o R^2 penalizado pelo número de variáveis. **F-statistic** é o resultado de um teste F de Fisher-Snedeco, para interpretar esse valor é necessário determinar um valor de alpha e usar uma tabela F, esse teste determina a significância estatística do nosso teste, **Prob(F-Statistic)** define a acurácia da hipótese nula. **Log-Likelihood** compara os valores dos coeficientes de cada variável na criação do modelo. **AIC** e **BIC** são usados para selecionar as características das variáveis.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     12 MacBook (Mid 2017, Gold)   R-squared:                       0.081
Model:                                     OLS   Adj. R-squared:                  0.042
Method:                          Least Squares   F-statistic:                     2.108
Date:                         Fri, 31 Mar 2023   Prob (F-statistic):              0.159
Time:                                 10:41:06   Log-Likelihood:                -50.564
No. Observations:                           26   AIC:                             105.1
Df Residuals:                               24   BIC:                             107.6
Df Model:                                    1
Covariance Type:                     nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           31.2066     20.835      1.498      0.147     -11.795      74.208
12 MacBook (Mid 2017, Gold)     -0.0195      0.013     -1.452      0.159      -0.047       0.008
==============================================================================
Omnibus:                       39.538   Durbin-Watson:                   1.513
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              129.539
Skew:                           2.893   Prob(JB):                     7.43e-29
Kurtosis:                      12.279   Cond. No.                     9.33e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.33e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Vamos discutir a parte inferior do sumário, **const** é o coeficiente linear, como discutido anteriormente a equação que rege a reta é dada por $y = a + bx$ onde a é o coeficiente linear, pode ser interpretado como o valor onde começa a nossa reta, ou ainda, onde o eixo vertical (y) é cortado, e então tem a nossa variável '**12 MacBook (Mid 2017, Gold)**' fechando os índices apresentados no nosso sumário vamos analisar as colunas a seguir. **Coef** para o nosso coeficiente linear é o valor dele, para a variável é a medida de como ela afeta o comportamento da reta, ou seja, é o nosso coeficiente angular o b da equação da reta. **std err** é desvio padrão, **t** é relacionado com quão preciso o coeficiente é medido. **P>|t|** é uma media importante que gera o p-valor, que mede a eficácia do modelo em medir a variável. **0,025 e 0,975** estabelecem as medidas de 95% dentro dos nosso dados, seguem a mesma definição clássica para outliers, que define outlier como dados fora de dois desvios padrões.

Omnibus é uma medida de normalidade que usa assimetria(skewness) e curtose, sendo 0 uma curva perfeitamente normal. **Prob(Omnibus)** é um teste estatístico que mede a normalidade da distribuição, o valor sendo 1 indica uma distribuição perfeitamente normal. **Skew** mede a assimetria da curva, com 0 zero temos um curva

perfeitamente simétrica. **Kurtosis** mede o quão agudo é o pico da nossa curva, altos valores indicam menos outliers. **Durbin-Watson** define a homocedaticidade ou uma distribuição uniforme dos erros de nossos dados, o ideal é ter valores entre 1 e 2. **Jarque-Bera(JB)** e **Prob(JB)** são métodos alternativos para medir os mesmos valores de Omnibus e Prob(Omnibus). E por fim **Cond. No.** Mede a sensibilidade do nosso modelo comparado com as mudanças nos nossos dados.