Regressão Linear e Testes de Hipótese

Nessa parte tentaremos encontrar uma equação que descreva a nota bruta dos estudantes do curso utilizando o método OLS(Ordinary least squares) para regressão linear

obs.: iremos adotar α_p=10%

O método OLS é uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados. É a forma de estimação mais amplamente utilizada na econometria. Consiste em um estimador que minimiza a soma dos quadrados dos resíduos da regressão, de forma a maximizar o grau de ajuste do modelo aos dados observados.
O modelo matemático para o método OLS é:

$$y = \alpha + \beta x_1 + \epsilon$$

onde:

- y = Variável que queremos descrever
- α = constante
- x_1 = Variável que usaremos para descrever y
- β = coeficiente da variável x_1
- ϵ = erro, representa a variação de y que o modelo não descreve

Inicialmente tentaremos descrever a Nota Bruta a partir das seguintes variáveis:
• Nota no componente especifico
• Tipo de escola que o aluno cursou no ensino médio
• Numero de pessoas que moram na mesma casa
• Renda familiar total
• Idade

```
resultado = sm.OLS(dados['NOTA BRUTA'],dados[['NOTA COMPONENTE ESPECÍFICO', 'ESCOLAEM', 'NPESSOASCASA',
'RENDAFAMILIA', 'IDADE']]).fit()
print(resultado.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             NOTA BRUTA   R-squared (uncentered):                   0.995
Model:                            OLS   Adj. R-squared (uncentered):              0.994
Method:                 Least Squares   F-statistic:                              1226.
Date:                Tue, 19 Nov 2019   Prob (F-statistic):                    1.22e-34
Time:                        08:42:16   Log-Likelihood:                         -95.352
No. Observations:                  36   AIC:                                      200.7
Df Residuals:                      31   BIC:                                      208.6
Df Model:                           5
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
NOTA COMPONENTE ESPECÍFICO     0.8831      0.041     21.792      0.000       0.800       0.966
ESCOLAEM                      -1.2331      1.382     -0.892      0.379      -4.052       1.586
NPESSOASCASA                   0.2770      0.345      0.803      0.428      -0.427       0.981
RENDAFAMILIA                  -0.0003      0.000     -0.947      0.351      -0.001       0.000
IDADE                          0.4032      0.099      4.069      0.000       0.201       0.605
==============================================================================
Omnibus:                        1.596   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.450   Jarque-Bera (JB):                1.226
Skew:                          -0.448   Prob(JB):                        0.542
Kurtosis:                       2.883   Cond. No.                     9.53e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.53e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Vamos começar comentando os dados da primeira tabela apresentada. As medidas mais importantes para nós neste momento são o R² ajustado(R-squared), a estatística de teste F (F-statisc) e o p-valor dessa estatística(Prob(F-statistic))

• O valor R² nos mostra o quanto esse modelo explica as variáveis, nesse caso $R^2 = 99.5\%$, ou seja, o modelo escolhido descreve muito bem a Nota Bruta
• a estatística de teste F e seu $p-valor < 0.001$ basicamente nos mostram que esse modelo é estatisticamente válido
analisando os coeficientes obtidos pela função **OLS** teremos a seguinte equação

$$y = 0.8831\,x_1 - 1.2331\,x_2 + 0.277\,x_3 - 0.0003\,x_4 + 0.4032\,x_5 + 1.867$$

onde:

- y = Nota Bruta
- x_1 = Nota Componente Específico
- x_2 = Tipo de escola que cursou no Ensino Médio
- x_3 = Número de pessoas que moram na mesma casa
- x_4 = Renda total familiar
- x_5 = Idade
- 1.867 = soma de todos os erros

porém, vemos alguns problemas com as variáveis escolhidas devido aos avisos dados na tabela. Tentaremos encontrar um novo modelo matemático que descreva a nota bruta removendo os dados imprecisos.

para sabermos quais dados válidos para a nova regressão iremos analisar a coluna $P>|t|$, pois quanto mais próximo seu valor de 0 mais relevante esse coeficiente é no modelo adotado,e como adotamos α_p =10%, iremos remover as variáveis que não atendem a esse critério (ESCOLAEM ,NPESSOASCASA , RENDAFAMILIA)

```
resultado = sm.OLS(dados['NOTA BRUTA'],dados[['NOTA COMPONENTE ESPECÍFICO','IDADE']]).fit()
print(resultado.summary())
```

```
                                 OLS Regression Results
================================================================================================
Dep. Variable:                NOTA BRUTA   R-squared (uncentered):                   0.995
Model:                            OLS   Adj. R-squared (uncentered):              0.994
Method:                 Least Squares   F-statistic:                              3109.
Date:                Tue, 19 Nov 2019   Prob (F-statistic):                    3.18e-39
Time:                        08:42:17   Log-Likelihood:                         -96.746
No. Observations:                  36   AIC:                                      197.5
Df Residuals:                      34   BIC:                                      200.7
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
NOTA COMPONENTE ESPECÍFICO     0.8799      0.040     22.258      0.000       0.800       0.960
IDADE                          0.3511      0.072      4.881      0.000       0.205       0.497
==============================================================================
Omnibus:                        0.719   Durbin-Watson:                   2.198
Prob(Omnibus):                  0.698   Jarque-Bera (JB):                0.739
Skew:                          -0.302   Prob(JB):                        0.691
Kurtosis:                       2.644   Cond. No.                         6.80
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

logo, analisando os novos coeficientes obtidos teremos a seguinte equação

$$y = 0.8799\,x_1 + 0.3511\,x_2 + 0.112$$

onde

- y = Nota Bruta
- x_1 = Nota Componente Específico
- x_2 = Idade
- 0.112 = soma dos erros

igualmente ao que fizemos na primeira análise iremos comentar os dados da tabela apresentada.

- R^2 =99.5%, ou seja, o modelo continua descrevendo muito bem a Nota Bruta
- temos a estatística de teste F e seu $p-valor < 0.001$ porém, nesse caso o valor de Prob(f) é ainda menor que o mostrado na primeira análise, o que nos indica que as variáveis usadas nessa nova regressão são ainda mais estatisticamente validos que os anteriores