Material Complementar - Correlação e Regressão

```
# Fit do modelo
X = sm.add_constant(selected_df.drop(columns=['price']))
y = selected_df['price']
model = sm.OLS(y, X).fit()

# Print regression summary
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.820
Model:                            OLS   Adj. R-squared:                  0.816
Method:                 Least Squares   F-statistic:                     227.4
Date:                Wed, 17 Apr 2024   Prob (F-statistic):           3.15e-73
Time:                        02:06:24   Log-Likelihood:                -1956.8
No. Observations:                 205   AIC:                             3924.
Df Residuals:                     200   BIC:                             3940.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.623e+04   1.28e+04     -3.610      0.000   -7.15e+04   -2.1e+04
carwidth     577.4872    224.477      2.573      0.011     134.841    1020.134
curbweight     2.1702      1.209      1.796      0.074      -0.213       4.554
enginesize    84.6046     12.587      6.722      0.000      59.784     109.425
horsepower    49.5919     10.556      4.698      0.000      28.778      70.406
==============================================================================
Omnibus:                       27.303   Durbin-Watson:                   0.768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.515
Skew:                           0.638   Prob(JB):                     5.35e-13
Kurtosis:                       5.234   Cond. No.                     1.40e+05
==============================================================================
```

1. X = sm.add_constant(selected_df.drop(columns=['price'])): Esta linha cria a matriz de features X para o modelo de regressão. selected_df.drop(columns=['price']) remove a coluna 'price' do dataframe selected_df, enquanto sm.add_constant adiciona uma coluna de uns à esquerda de X, necessária para o termo constante na regressão.
2. y = selected_df['price']: Esta linha define a variável dependente y como sendo a coluna 'price' do dataframe selected_df.
3. model = sm.OLS(y,X).fit(): Aqui, o modelo de regressão linear é ajustado aos dados usando a função sm.OLS da biblioteca StatsModels. y é a variável dependente e X é a matriz de features. O método .fit() é chamado para ajustar o modelo aos dados.
4. print(model.summary()): Finalmente, esta linha imprime um resumo dos resultados da regressão, incluindo estatísticas como coeficientes, erros padrão, valores p, estatísticas t, R-quadrado, etc. Isso fornece uma visão geral do desempenho e significância do modelo ajustado.

Dos resultados:

1. **Método e Tipo de Covariância:**
    - "Method: least squares": Isso indica que o método usado para ajustar o modelo foi o método dos mínimos quadrados, que é comumente usado em regressões lineares.
    - "Covariance type: nonrobust": Isso indica que a estimativa da matriz de covariância não é robusta a outliers ou erros não gaussianos nos dados.
2. **R-quadrado e R-quadrado ajustado:**
    - "R-squared: 0.82": O coeficiente de determinação (R-quadrado) é 0.82, o que significa que aproximadamente 82% da variabilidade na variável dependente (price) é explicada pelas variáveis independentes no modelo.
    - "Adjusted R-squared: 0.816": O R-quadrado ajustado leva em consideração o número de preditores no modelo e é uma versão mais conservadora do R-quadrado, penalizando modelos com muitos preditores que não contribuem significativamente para explicar a variabilidade da variável dependente.
3. **Estatísticas F e Probabilidade F:**
    - "F-statistic: 227.4": A estatística F é usada para testar a significância global do modelo de regressão. Neste caso, o valor alto da estatística F (227.4) sugere que o modelo como um todo é estatisticamente significativo.
    - "Prob (F-statistic): 3.15e-73": Esta é a probabilidade associada à estatística F. É um valor muito baixo (3.15e-73 significa 3.15 vezes 10 elevado a -73), o que indica que a probabilidade de obter uma estatística F tão alta por acaso é extremamente baixa, reforçando a significância global do modelo.
4. **Log Likelihood, AIC e BIC:**
    - "Log likelihood: -1956.8": A log-verossimilhança é uma medida da adequação do modelo aos dados. Quanto maior o valor absoluto da log-verossimilhança negativa (como é o caso aqui), melhor o ajuste do modelo aos dados.
    - "AIC: 3924" (Akaike Information Criterion): O AIC é um critério de seleção de modelo que penaliza modelos mais complexos. Quanto menor o valor do AIC, melhor é o ajuste do modelo.
    - "BIC: 3940" (Bayesian Information Criterion): Assim como o AIC, o BIC é um critério de seleção de modelo que penaliza a complexidade. Um valor menor indica um modelo melhor ajustado.
5. **Modelo e Variable Dependente**
    - “Dep. Variable : price”: lista qual a variable dependente usada no modelo.
    - “Model: OLS”: explicita qual o modelo usado para o calculo dessa regressao

Em resumo, os resultados indicam que o modelo de regressão linear ajustado parece ser estatisticamente significativo, explicando uma porcentagem substancial da variabilidade na variável dependente e apresentando um bom ajuste aos dados.
A seção que menciona as colunas "const", "std err", "t", "P>|t|", "[0.025 0.975]" é a parte do resumo da regressão que mostra os coeficientes estimados para cada variável independente no modelo.

1. **const:**
    - A coluna "const" representa o coeficiente do termo constante (intercepto) da regressão linear. Este valor é a estimativa do valor médio da variável dependente quando todas as variáveis independentes são zero.
2. **std err (Erro Padrão):**

- O "std err" é o erro padrão do coeficiente estimado. Ele indica a precisão da estimativa do coeficiente. Quanto menor o erro padrão, mais confiável é a estimativa do coeficiente.

3. **t (Estatística t):**
   - A estatística "t" é calculada dividindo o coeficiente pelo seu erro padrão. Ela indica a significância do coeficiente. Quanto maior o valor absoluto da estatística t, mais significativo é o coeficiente.
4. **P>|t| (Valor p):**
   - O valor "P>|t|" (valor p) é a probabilidade de observar uma estatística t tão extrema (ou mais extrema) sob a hipótese nula de que o coeficiente é zero (ou seja, de que a variável independente não tem efeito sobre a variável dependente).
   - Se o valor p for menor que um nível de significância escolhido (como 0.05), geralmente consideramos o coeficiente como estatisticamente significativo.
5. **[0.025 0.975] (Intervalo de Confiança):**
   - O intervalo "[0.025 0.975]" é o intervalo de confiança para o coeficiente estimado. Ele fornece uma faixa dentro da qual acredita-se que o verdadeiro valor do coeficiente esteja com uma certa probabilidade (geralmente 95%).

Quanto a ultima sessão, com algumas estatísticas adicionais:

1. **Omnibus:**
   - O teste Omnibus é uma medida global da normalidade dos resíduos do modelo de regressão. Se o valor do Omnibus for significativamente diferente de zero, isso indica que os resíduos não estão distribuídos normalmente. Um valor baixo pode sugerir que os pressupostos do modelo (como a normalidade dos resíduos) podem não ser atendidos.
2. **Prob(Omnibus):**
   - Este é o valor p associado ao teste Omnibus. Ele indica a probabilidade de obter um valor de Omnibus tão extremo ou mais extremo do que o observado, assumindo que os resíduos são normalmente distribuídos. Um valor p baixo sugere que os resíduos não são normalmente distribuídos.
3. **Skew (Assimetria):**
   - A assimetria (skewness) mede a falta de simetria na distribuição dos resíduos. Um valor de assimetria diferente de zero indica uma distribuição assimétrica. Um valor positivo indica uma cauda mais longa à direita, enquanto um valor negativo indica uma cauda mais longa à esquerda.
4. **Kurtosis (Curtose):**
   - A curtose (kurtosis) mede a "cauda" de uma distribuição de resíduos. Valores de curtose maiores que zero indicam uma cauda mais pesada (distribuição mais concentrada ao redor da média, com caudas mais longas). Valores menores que zero indicam uma cauda mais leve.
5. **Durbin-Watson:**
   - O teste de Durbin-Watson é usado para detectar a presença de autocorrelação nos resíduos (ou seja, se os resíduos estão correlacionados no tempo ou em ordem de observação). O valor ideal do teste de Durbin-Watson está entre 1 e 2; valores menores que 1 sugerem autocorrelação positiva, enquanto valores maiores que 2 sugerem autocorrelação negativa.

6. **Jarque-Bera (JB) e Prob(Jarque-Bera):**
    - O teste de Jarque-Bera é outra medida de normalidade dos resíduos. Valores altos de JB indicam desvios da normalidade. O valor p associado ao teste de Jarque-Bera indica a probabilidade de obter um valor de JB tão extremo ou mais extremo do que o observado, assumindo que os resíduos são normalmente distribuídos.
7. **Cond No (Número de Condição):**
    - O número de condição é uma medida da multicolinearidade no modelo. Valores altos de número de condição indicam multicolinearidade, o que pode prejudicar a precisão das estimativas dos coeficientes.

Essas estatísticas são úteis para avaliar a adequação do modelo de regressão e identificar possíveis problemas, como a falta de normalidade dos resíduos, autocorrelação, assimetria, curtose, multicolinearidade, entre outros. Valores significativos em algumas dessas estatísticas podem indicar a necessidade de investigação adicional ou ajustes no modelo.