**Análise de Regressão Múltipla**[**¶**](http://localhost:8888/notebooks/Documents/datascience_course/2-analise-explorat%C3%B3ria-inter/03-regress%C3%B5es%20m%C3%BAltiplas/reg_mult_rascunho.ipynb#Análise-de-Regressão-Múltipla)

**Introdução**[**¶**](http://localhost:8888/notebooks/Documents/datascience_course/2-analise-explorat%C3%B3ria-inter/03-regress%C3%B5es%20m%C3%BAltiplas/reg_mult_rascunho.ipynb#Introdução)

A regressão múltipla é uma extensão da regressão linear simples. É usado quando queremos predizer o valor de uma variável com base no valor de duas ou mais outras variáveis. A variável que queremos prever é chamada de variável dependente (ou, às vezes, variável de resultado, meta ou critério). As variáveis ​​que estamos usando para prever o valor da variável dependente são chamadas de variáveis ​​independentes (ou às vezes, as variáveis ​​preditoras, explicativas ou regressivas).

Por exemplo, é possível usar a regressão múltipla para entender se o desempenho do exame pode ser previsto com base no tempo de revisão, na ansiedade do teste, na frequência de palestras e no sexo. Alternativamente, seria possível usar a regressão múltipla para entender se o consumo diário de cigarros pode ser previsto com base na duração do tabagismo, idade quando começou a fumar, tipo de fumante, renda e sexo.

A regressão múltipla também permite determinar o ajuste geral (variância explicada) do modelo e a contribuição relativa de cada um dos preditores para a variância total explicada. Por exemplo, se existe a necessidade de saber quanto da variação no desempenho do exame pode ser explicado pelo tempo de revisão, teste de ansiedade, freqüência de palestras e gênero "como um todo", mas também a "contribuição relativa" de cada variável independente para explicar variância.

Este texto demonstra como realizar a regressão múltipla, bem como interpretar e relatar os resultados deste teste. No entanto, antes de apresentar esse procedimento, você precisa entender as diferentes suposições que seus dados devem atender para que a regressão múltipla forneça um resultado válido. Nós discutimos essas suposições a seguir.

**Premissas**[**¶**](http://localhost:8888/notebooks/Documents/datascience_course/2-analise-explorat%C3%B3ria-inter/03-regress%C3%B5es%20m%C3%BAltiplas/reg_mult_rascunho.ipynb#Premissas)

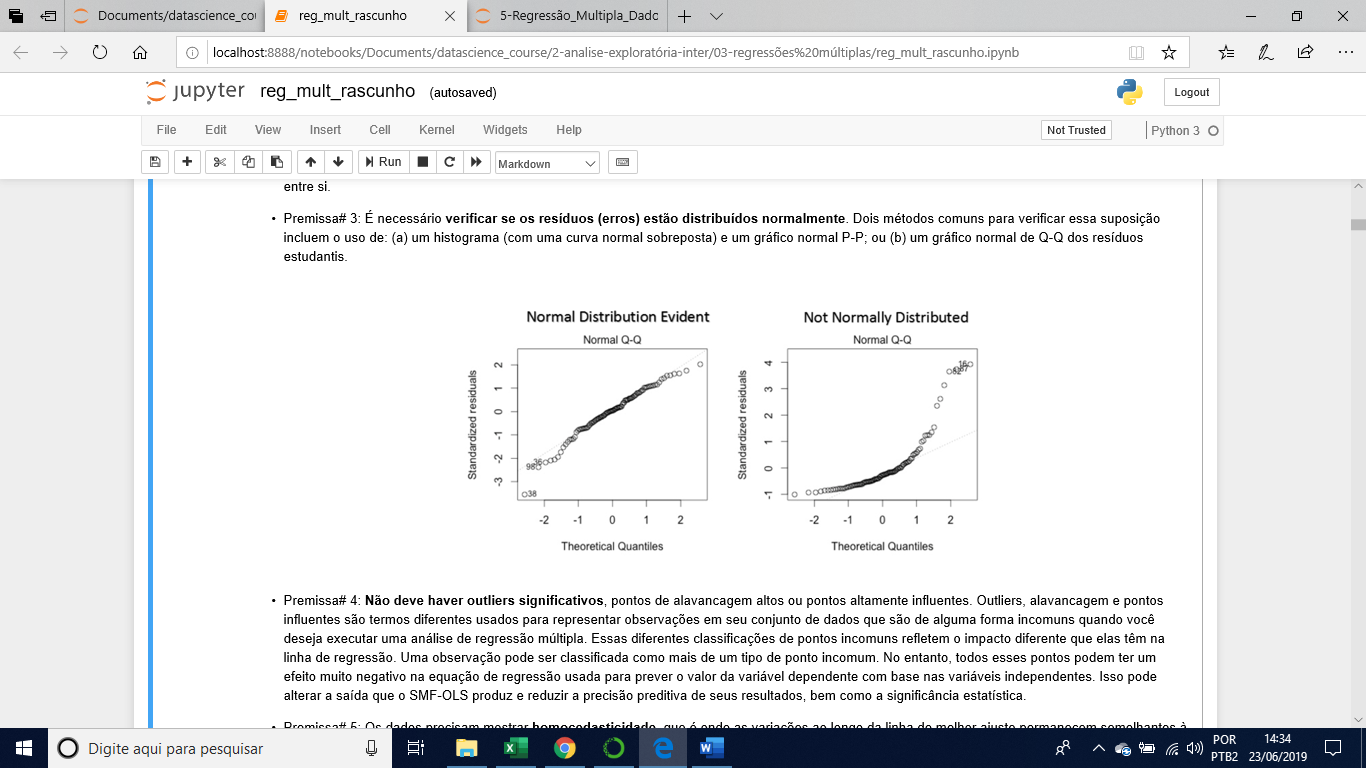
Quando se escolhe analisar dados usando a regressão múltipla, parte do processo envolve a verificação para garantia de que os dados possam realmente ser analisados ​​usando a regressão múltipla. É preciso fazer isso porque só é apropriado usar a regressão múltipla se os dados "passarem" por oito hipóteses que são necessárias para a regressão múltipla forneça resultados válidos. Na prática, a verificação desses oito pressupostos apenas adiciona um pouco mais de tempo à análise, exigindo alguns testes e vizualizações dos dados, porém não representa uma tarefa difícil.

Antes de apresentar as oito hipóteses sobre a análise dos dados, é importante ressaltar que é possíver haver uma violação de uma ou mais dessas suposições (ou seja, suposições que não sejam atendidas). Isso não é incomum quando se trabalha com dados do mundo real em vez de exemplos de livros didáticos, que muitas vezes mostram apenas como realizar a regressão múltipla quando tudo está em estado perfeito. No entanto, mesmo quando os dados falham em certas suposições, geralmente há uma solução para superar isso. Primeiro, vamos dar uma olhada nessas oito suposições:

A variável dependente deve ser medida em uma escala contínua (isto é, uma variável de intervalo ou razão). Exemplos de variáveis ​​que atendem a esse critério incluem tempo de revisão (medido em horas), inteligência (medida pelo índice de QI), desempenho do exame (medido de 0 a 100), peso (medido em kg) e assim por diante. Se a variável dependente for medida em uma escala ordinal, será necessário realizar uma regressão ordinal em vez de uma regressão múltipla. Exemplos de variáveis ​​ordinais incluem itens Likert (por exemplo, uma escala de 7 pontos de "concordo totalmente" até "discordo totalmente"), entre outras formas de classificação de categorias (por exemplo, uma escala de 3 pontos explicando quanto um cliente gostou de um produto , variando de "Não muito" para "Sim, muito".

Além disso, é necessário ter duas ou mais variáveis ​​independentes, que podem ser contínuas (isto é, uma variável de intervalo ou razão) ou categóricas (ou seja, uma variável ordinal ou nominal). Exemplos de variáveis ​​nominais incluem gênero (por exemplo, 2 grupos: masculino e feminino), etnia (por exemplo, 3 grupos: caucasianos, afro-americanos e hispânicos), nível de atividade física (por exemplo, 4 grupos: sedentários, baixos, moderados e altos) profissão (por exemplo, 5 grupos: cirurgião, médico, enfermeiro, dentista, terapeuta) e assim por diante. Se uma de suas variáveis ​​independentes for dicotômica e considerada uma variável moderadora, talvez seja necessário executar uma análise de moderador Dicotômica.

* Premissa# 1: É preciso haver uma **relação linear** entre (a) a variável dependente e cada uma de suas variáveis ​​independentes, e (b) a variável dependente e as variáveis ​​independentes coletivamente. Embora haja várias maneiras de verificar esses relacionamentos lineares, é sugerida a criação de gráficos de dispersão e gráficos de regressão parcial usando Matplotlib Pyplot, Seaborn Jointplot ou Pairplot, e em seguida, inspecionar visualmente esses gráficos de dispersão e gráficos de regressão parcial para verificar a linearidade. Se o relacionamento exibido em seus gráficos de dispersão e gráficos de regressão parcial não forem lineares, é preciso executar uma análise de regressão não linear ou "transformar" seus dados. Dentre as soluções mais utilizadas para este tipo de situação estão: (a) criar gráficos de dispersão e gráficos de regressão parcial para verificar a linearidade ao realizar a regressão múltipla; (b) interpretar diferentes resultados de plotagem de dispersão e de regressão parcial; e (c) transformar seus dados caso não exista relações lineares entre suas variáveis. (Isso dá Bestas constantes, e taxas constantes).
* Premissa# 2: É necessário ter **independência de observações** (ou seja, independência de resíduos), que pode ser verificado usando a estatística Durbin-Watson, que é um teste simples para ser realizado usando o pacote SMF (Stats Model Function) OLS (Ordinary Least Squares), cujo resultado estará disponível juntamente com o resultado da regressão. Isto acontece pois os dados não devem mostrar multicolinearidade, que ocorre quando se tem duas ou mais variáveis ​​independentes altamente correlacionadas entre si. Isso leva a problemas com o entendimento de qual variável independente contribui para a variância explicada na variável dependente, assim como questões técnicas no cálculo de um modelo de regressão múltipla. Como boas práticas de verificação, é executado um código de cálculo das correlações múltiplas para verificação dos valores de correlação de todas as variáveis entre si.
* Premissa# 3: É necessário **verificar se os resíduos (erros) estão distribuídos normalmente**. Dois métodos comuns para verificar essa suposição incluem o uso de: (a) um histograma (com uma curva normal sobreposta) e um gráfico normal P-P; ou (b) um gráfico normal de Q-Q dos resíduos estudantis.



* Premissa# 4: **Não deve haver outliers significativos**, pontos de alavancagem altos ou pontos altamente influentes. Outliers, alavancagem e pontos influentes são termos diferentes usados ​​para representar observações em seu conjunto de dados que são de alguma forma incomuns quando você deseja executar uma análise de regressão múltipla. Essas diferentes classificações de pontos incomuns refletem o impacto diferente que elas têm na linha de regressão. Uma observação pode ser classificada como mais de um tipo de ponto incomum. No entanto, todos esses pontos podem ter um efeito muito negativo na equação de regressão usada para prever o valor da variável dependente com base nas variáveis ​​independentes. Isso pode alterar a saída que o SMF-OLS produz e reduzir a precisão preditiva de seus resultados, bem como a significância estatística.
* Premissa# 5: Os dados precisam mostrar **homocedasticidade**, que é onde as variações ao longo da linha de melhor ajuste permanecem semelhantes à medida que se move ao longo da linha. Ao analisar seus dados, é preciso plotar os resíduos em relação aos valores previstos não padronizados em um gráfico scatter plot, e verificar se não existe nenhum padrão definido entre os resíduos e os valores previstos.

