Aula 9: Transformação de Variáveis e Interações

**Lista de Exercícios**

31 de outubro & 02 de novembro de 2022

Para realizar esta lista de exercícios, considere a base da dados que está no *Moodle*. Esta é uma base com dados reais, com os resultados das eleições nacionais de 2010 e 2014 – a mesma da semana passada.

Para garantir que o exercício será completado satisfatoriamente, instale as seguintes bibliotecas e carregue os seguintes pacotes:

install.packages("sjPlot")

install.packages("sjmisc")

library(sjPlot)

library(sjmisc)

library(tidyverse)

O banco nos permite responder a questão sobre o que explicou o voto no PT no ano de 2014 (PT\_2014). A unidade de análise é a votação por município. Então, 5572 observações. Há, como variáveis explicativas, o voto no partido em 2010 (PT\_2010); a população das cidades, tanto em nível (pop2014), quanto em logaritmo (ln\_pop\_2014); a porcentagem de famílias da cidade que recebem o bolsa família em 2014 (alcance\_PBF\_2014) e uma *dummy* que indica se a cidade está na região Nordeste do país ou não (Nordeste).

Inicialmente, precisamos avaliar se as variáveis explicativas se adequariam a um modelo linear em relação á variável resposta (PT\_2014). Para isto, vamos lançar mão de gráficos que nos ajudarão a tomar essa decisão.

Usando o comando a seguir, construa gráficos de dispersão para cada um dos pares de variáveis, alterando a variável x (marcada em vermelho).

**ggplot(X2010\_2014, aes(x =** **PT\_2010**, **y = PT\_2014)) + geom\_point()**

1. Discuta a adequação do modelo linear para cada variável. Determine, assim, as variáveis que entrarão no modelo e sua especificação.
2. Se rodássemos um modelo bivariado apenas com a variável transformada, como interpretaríamos o coeficiente deste modelo?

Vamos agora rodar o modelo com a variável transformada mais todas as demais variáveis apresentadas.

1. Apresente os resultados da regressão. Interprete os parâmetros estimados.

De acordo com a visão convencional, inclusive do senso comum, o bolsa família teve muita influência na eleição e ainda mais nas cidades nordestina. Lá, os efeitos se reforçaram, dando ainda mais vantagem para o Partido dos Trabalhadores.

1. A partir do modelo rodado anteriormente, altere a especificação e forma a contemplar este entendimento. Apresente os resultados do modelo.
2. Construa um gráfico que mostre os efeitos marginais destes novos termos introduzidos. Para isto, utilize o comando a seguir:

**plot\_model(*modelo*, type = "pred", terms =c("variável contínua", "variável *dummy*"))**

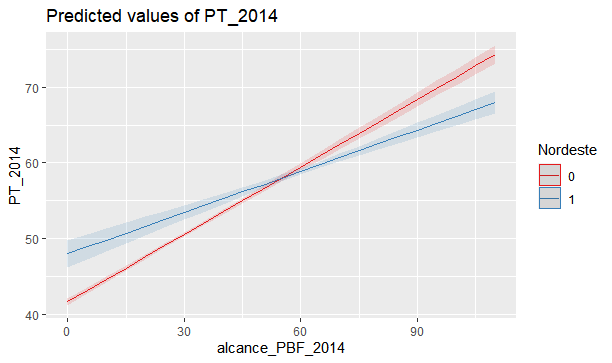
em que *modelo* é o rótulo do objeto onde você registrou os resultados da regressão; *variável contínua* é a variável explicativa contínua e a *variável dummy* é a variável explicativa *dummy* utilizadas no modelo.

1. Interprete seu gráfico, comparando-o com os resultados da regressão que você obteve no item 4).

Para realizar esta segunda discussão, não haverá exercício propriamente dito. Vamos testar as críticas de Hainmueller, Mummolo e Xu (2019), implementando seus comandos sugeridos, em relação à base que utilizamos aula passada. A ideia é reavaliarmos se nossos resultados obtidos se sustentam com a crítica feita por eles a trabalhos que utilizam modelos interativos.

Suponha, então, que tenhamos o mesmo modelo da aula passada, que é algo um pouco fora do usual quando se pensa um experimento: a cidade estar no Nordeste é o tratamento e a cobertura do Bolsa Família, a variável moderadora. Isto não é preciso do ponto de vista metodológico porque o tratamento precisaria ser aleatório e no caso não é. Fazemos isso pelo exercício apenas, já que precisaríamos de uma *dummy* para facilitar a noção de tratamento. Assim, nosso “experimento” aqui é: se alocássemos uma cidade no Nordeste (tratamento), teria ela apresentado uma votação maior para o PT do que uma cidade que está fora do Nordeste (controle)?

Partimos, então, do resultado que produzimos na aula passada. Sua representação está abaixo:



1. Observando este gráfico, é possível sabermos se as críticas apontadas por Hainmueller, Mummolo e Xu (2019) se aplicam a este caso? Aponte as duas críticas.

As duas críticas apontadas pelos autores são: se a relação entre as variáveis não é linear, um modelo linear pode ser usado apenas como uma aproximação linear do que seria o modelo real (p.164). A segunda crítica é que é necessário atentar-se se há *common support* nos dados quando montamos um modelo, isto né, no caso do nosso exemplo, se existe um número suficiente de casos em que Z=0 e Z=1, se não a conclusão que tiramos do nosso modelo é mera extrapolação.

A fim de nos certificarmos de se as críticas se aplicam, adotamos a primeira das sugestões feitas pelos autores e sobrepusemos três gráficos: o de dispersão entre alcance do BF (eixo X) e a votação no PT (eixo Y), a regressão linear (*Fitted values*) e o gráfico de ajuste ponderado (*Loess regression*). No gráfico abaixo podemos ver os resultados destes gráficos, quando as cidades não são do Nordeste (*Nordeste = 0* – gráfico a esquerda) e quando as cidades são do Nordeste (*Nordeste = 1* – gráfico a direita):



1. Observando o gráfico acima, quais são suas posições a respeito das críticas dos autores apontadas na resposta 1?

Ainda seguindo a proposição dos autores, vamos executar a primeira de suas proposições que é a de regressão linear baseada em intervalos (*bins*). Através desta regressão, teremos condições de confirmar se a relação é mesmo linear ou não. O resultado da regressão é dividido em 3 valores representativos de cada tercil dos dados, e chamado de Baixo (L), Médio (M), Alto (H). É possível também considerar o suporte da variável moderadora. O gráfico abaixo apresenta os resultados:



1. Em termos da linearidade do efeito e da distribuição do suporte, quais conclusões podemos tirar a partir do gráfico acima? Além disso, a inclinação da reta é condizente com o resultado obtido no primeiro gráfico apresentado aqui (encontrado na aula passada)? Por que?

Vamos, então, partir para uma avaliação semi-paramétrica. Esta nos permitirá investigar qualquer suposta não linearidade dos efeitos do município estar no Nordeste (tratamento), moderado pelo alcance do Bolsa Família. O gráfico abaixo apresenta um resultado obtido usando o mesmo comando proposto pelos autores. É um modelo via *kernel*. O *bin* aqui é de 10,6, ajuste ótimo calculado pelo próprio modelo.



1. Qual a conclusão podemos tirar do efeito do nosso tratamento sobre a votação para o PT, em termos de linearidade e de suporte da variável moderadora? Em relação ao nosso resultado obtido na aula passada, ele se sustenta? Por que?