**Desafio Pé de Meia**

**VISÃO GERAL**

Avaliar a efetividade do programa **Pé de Meia** em incentivar a permanência e conclusão do ensino médio entre alunos de escolas públicas que pertencem a um grupo vulnerável (idade entre 14 e 24 anos e faixa de renda abaixo de 1/2 salário-mínimo). Além disso:

1. **Caracterizar o público-alvo**: Quantificar e entender o perfil socioeconômico dos alunos elegíveis, usando dados da PNAD e do CadÚnico, e observar aspectos como estrutura familiar, composição do domicílio e escolaridade dos pais.
2. **Analisar padrões de abandono e evasão escolar**: Diferenciar entre alunos que interrompem e aqueles que continuam seus estudos, usando dados de frequência e indicadores de escolaridade ao longo dos trimestres. Você pretende utilizar modelos multivariados (como probit e logit) para relacionar abandono e evasão com variáveis socioeconômicas e características do programa.
3. **Simular o impacto do programa**: Estimar custos e projetar cenários para o programa, comparando o cenário atual com um cenário ideal em que todos os alunos atingem 80% de frequência e evitam a evasão. Isso inclui também a análise dos incentivos oferecidos (matrícula, frequência, conclusão e ENEM).
4. **Estimação da efetividade**: Analisar a efetividade do programa em termos de retenção escolar e estímulo à conclusão do ensino médio, considerando a comparação de dados entre trimestres e o uso de variáveis como renda per capita e escolaridade parental para estimar os resultados e projetar melhorias.

**MODELOS ECONOMÉTRICOS**

Pode-se considerar os seguintes modelos econométricos, cada um atendendo a uma necessidade específica da análise:

**1. Modelos de Regressão Logística (Logit) e Probit**

* **Objetivo**: Analisar a probabilidade de abandono escolar e evasão.
* **Aplicação**: Você pode usar esses modelos para estimar a probabilidade de um aluno permanecer no ensino médio ou evadir-se, considerando características como idade, escolaridade dos pais, renda familiar per capita, e participação no programa **Pé de Meia**.
* **Interpretação**: Os coeficientes desses modelos indicarão o efeito marginal das variáveis explicativas (por exemplo, renda e escolaridade) na probabilidade de conclusão do ensino médio.

**2. Modelos Heckman (Heckit)**

* **Objetivo**: Corrigir o viés de seleção em relação ao grupo elegível e à adesão ao programa.
* **Aplicação**: Como o programa **Pé de Meia** é destinado a alunos em condição de vulnerabilidade, você pode usar o modelo Heckman para controlar o viés de seleção, avaliando como a participação (ou elegibilidade) ao programa influencia a retenção e conclusão do ensino médio.
* **Interpretação**: A primeira etapa do Heckman estima a probabilidade de participação no programa, enquanto a segunda etapa avalia o efeito do programa nos alunos que realmente participam. Isso ajuda a identificar se o programa está causando retenção e conclusão ou se outros fatores correlacionados à seleção estão influenciando.

**3. Modelos de Diferenças em Diferenças (DiD)**

* **Objetivo**: Comparar a mudança na taxa de abandono antes e depois da implementação do programa.
* **Aplicação**: Se o programa **Pé de Meia** foi implementado em uma data específica, você pode comparar os alunos que participaram antes e depois do programa com um grupo controle que não teve acesso. Este modelo ajuda a identificar o efeito causal do programa sobre o abandono e a conclusão.
* **Interpretação**: A variação da taxa de abandono entre os grupos antes e depois do programa fornece uma medida da efetividade do programa.

**4. Modelos de Sobrevivência (Análise de Tempo de Permanência)**

* **Objetivo**: Estimar o tempo até a evasão ou conclusão do ensino médio.
* **Aplicação**: Utilizando os dados de múltiplos trimestres da PNAD, este modelo permite analisar quanto tempo os alunos permanecem no ensino médio antes de sair ou concluir. É útil para examinar se o programa **Pé de Meia** prolonga o tempo de permanência.
* **Interpretação**: A análise indica o efeito do programa sobre o “tempo de sobrevivência” no ensino médio, podendo medir se alunos participantes do programa têm maior probabilidade de continuar até a conclusão.

**5. Modelos de Propensão (Propensity Score Matching)**

* **Objetivo**: Comparar alunos participantes e não participantes do programa com perfis semelhantes.
* **Aplicação**: Para identificar o impacto do programa em alunos com características similares, você pode usar o emparelhamento por escores de propensão, criando grupos comparáveis com e sem acesso ao **Pé de Meia**.
* **Interpretação**: Esse método ajuda a isolar o efeito do programa, garantindo que as diferenças de conclusão e abandono não sejam atribuídas a variáveis como renda ou escolaridade dos pais.

**6. Modelos de Dados em Painel**

* **Objetivo**: Analisar a trajetória dos mesmos indivíduos ao longo dos trimestres.
* **Aplicação**: Os dados trimestrais permitem estimar um modelo de painel para controlar variáveis inobserváveis e observar mudanças individuais nos alunos. Com isso, você pode identificar os fatores que impactam a permanência no ensino médio ao longo do tempo.
* **Interpretação**: Os efeitos fixos e aleatórios controlam fatores individuais constantes, ajudando a medir a influência do programa sobre a evolução dos alunos.

Esses modelos em conjunto oferecem uma análise robusta sobre a efetividade do programa **Pé de Meia**, permitindo controlar por vieses, observar impactos temporais e explorar tanto efeitos imediatos quanto de longo prazo na educação dos alunos.

**PONTOS DE ALINHAMENTO E REFINAMENTO**

* **Objetivo 1 (Caracterização do Público-Alvo)**: Os dados carregados e transformados, como idade, renda familiar, e escolaridade dos pais, estão prontos para uma análise de perfil.
* **Objetivo 2 (Análise de Padrões de Abandono e Evasão)**: O próximo passo é garantir que as variáveis de escolaridade e trabalho estejam preparadas para análises logit/probit, que podem modelar a probabilidade de evasão escolar com base nas variáveis socioeconômicas.
* **Objetivo 3 (Simulação de Impacto)**: Para a simulação, você poderá expandir as variáveis de frequência escolar e os incentivos. Considere criar dummy variables para simular cenários, como frequências abaixo ou acima de 80%.
* **Objetivo 4 (Estimação de Efetividade)**: A continuidade dos dados entre trimestres, incluindo renda per capita ajustada, está bem definida, mas a análise longitudinal exigirá o cálculo de métricas trimestrais.

**SOBRE A APLICAÇÃO DOS MODELOS**

Você pode realizar a análise de padrões de abandono e evasão com o dataframe amostral da PNAD (pnad), desde que incorpore os pesos de amostragem e a estrutura de estratificação e clusterização, que são essenciais para que os resultados dos modelos probit, logit ou Heckman possam ser interpretados para a população.

**Passos para Garantir a Representatividade**

1. **Aplicação dos Pesos de Amostragem**:
   * Os pesos da PNAD (como V1031 e V1032) representam a expansão amostral para a população. Ao usá-los, os modelos probit e logit ajustados no dataframe amostral podem ser interpretados em relação à população, corrigindo distorções que surgiriam em uma amostra simples não ponderada.
2. **Estrutura de Estratificação e Clusterização**:
   * Incorporar a estrutura de estratos e clusters no design amostral também é importante para evitar vieses. Com o pacote survey, você pode definir um design amostral que incorpore essa estrutura, o que ajusta as estimativas de variância e evita interpretações incorretas sobre a população.
3. **Modelagem com Ajustes Ponderados**:
   * Ao configurar o design, você pode aplicar modelos probit e logit ponderados. Usando a estrutura de svyglm() do survey, por exemplo, os coeficientes do modelo podem ser estimados de forma robusta, e a interpretação dos coeficientes no contexto da população é adequada.

**Heckman (ou Heckit)**

* No caso do modelo Heckman, que considera um processo de seleção, o uso de pesos amostrais torna-se mais complexo. Para estimar corretamente com pesos e a estrutura de estratificação, pode ser necessário o uso de pacotes que permitam a combinação de pesos com seleção, como o pacote sampleSelection. Alternativamente, uma abordagem bayesiana pode ajudar a lidar com amostras complexas em cenários com seleção.

**Conclusão**

A análise de abandono e evasão com a amostra da PNAD é viável para interpretar a população, desde que:

* Os pesos e a estrutura de estratificação e clusterização sejam devidamente aplicados.
* Os modelos sejam ajustados considerando essas especificações para garantir estimativas robustas e representativas.

Assim, você poderá inferir sobre toda a população com base nos dados amostrais, o que é ideal para o propósito do programa Pé de Meia.

**COLUNAS df pnad2**

1. **Ano**: Ano da coleta dos dados.
2. **Trimestre**: Trimestre do ano em que a coleta foi realizada.
3. **UF**: Unidade da Federação (estado brasileiro).
4. **RM\_RIDE**: Região Metropolitana ou Região Integrada de Desenvolvimento Econômico (indica se o domicílio está em uma dessas regiões).
5. **UPA**: Unidade Primária de Amostragem, uma identificação para organizar a amostra.
6. **Estrato**: Estrato amostral, utilizado para organização da amostra dentro da pesquisa.
7. **V1008**: Número da unidade domiciliar na amostra.
8. **V1014**: Tipo de domicílio, como casa, apartamento, etc.
9. **V1022**: Tipo de área, classificada como urbana ou rural.
10. **V1023**: Localização do domicílio dentro da UF, como capital, interior, etc.
11. **V2001**: Número de moradores no domicílio.
12. **V2003**: Código de parentesco com o responsável do domicílio.
13. **V2005**: Relação do morador com o responsável pelo domicílio (ex. cônjuge, filho).
14. **V2007**: Sexo do morador (masculino/feminino).
15. **V20081:** Mês de nascimento do morador.
16. **V20082**: Ano de nascimento do morador.
17. **V2009**: Idade do morador.
18. **V2010**: Cor ou raça do morador.
19. **V3002**: Indicador de frequência escolar (se o morador está frequentando a escola).
20. **V3002A**: Tipo de rede escolar, pública ou privada.
21. **V3003A**: Nível ou modalidade de ensino (ensino fundamental, médio, etc.).
22. **V3006**: Série ou ano escolar atual do morador.
23. **V3009A**: Modalidade de ensino atual.
24. **V4001**: Indicador de trabalho (se o morador trabalha).
25. **V4009**: Número de empregos.
26. **V5001A**: Indicador de recebimento de benefício do Programa Bolsa Família.
27. **V5001A2**: Valor recebido pelo Bolsa Família.
28. **V5002A**: Indicador de recebimento de benefício do Benefício de Prestação Continuada (BPC).
29. **V5002A2**: Valor recebido pelo BPC.
30. **V5003A**: Indicador de recebimento de outro benefício social.
31. **V5003A2**: Valor recebido por outro benefício social.
32. **VD2002**: Posição no domicílio (responsável, cônjuge, filho, etc.).
33. **VD2003**: Número de pessoas que contribuem com renda no domicílio.
34. **VD2004**: Tipo de composição familiar (nuclear, extensa, etc.).
35. **VD3006**: Nível de instrução ou escolaridade.
36. **VD4002**: Situação de ocupação (empregado, desocupado).
37. **VD4008**: Tipo de ocupação, como empregado no setor privado, público, etc.
38. **VD4009**: Tipo de vínculo empregatício (com ou sem carteira de trabalho).
39. **VD4010**: Setor de atividade econômica.
40. **VD4011**: Grupo ocupacional.
41. **VD4013**: Jornada semanal de trabalho em horas.
42. **VD4016**: Rendimento do trabalho principal.
43. **VD4017**: Outros rendimentos do trabalho.
44. **VD4047**: Rendimento mensal domiciliar total.
45. **ID\_DOMICILIO**: Identificador único do domicílio.
46. **GR**: Grande Região (Norte, Nordeste, Sudeste, Sul, Centro-Oeste).
47. **V2001T**: Total de moradores no domicílio.
48. **VD4016AP**: Rendimento do trabalho principal ajustado pela inflação.
49. **VD4017AP**: Outros rendimentos do trabalho ajustados pela inflação.
50. **VD4047AP**: Rendimento mensal domiciliar ajustado pela inflação.
51. **V2001R**: Número de moradores no domicílio (restrito a uma definição específica da PNADC).
52. **VD4016R**: Rendimento do trabalho principal restrito.
53. **VD4017R**: Outros rendimentos do trabalho restrito.
54. **VD4047R**: Rendimento mensal domiciliar restrito.
55. **RD**: Rendimento domiciliar total.
56. **RDPC**: Rendimento domiciliar per capita.
57. **RDPC\_categoria**: Categoria de rendimento domiciliar per capita (faixa de renda).

# Definir as variáveis de interesse

variaveis <- c(

**# Localização**

'UF', # Unidade da Federação

'V1022', # Situação do domicílio (rural ou urbana)

'V1023', # Tipo de área (Capital, RM, RIDE e UF)

'RM\_RIDE', # Região Metropolitana (rm de capitais brasileiras)

**# Condições do domicílio**

'V1008', # Número de seleção do domicílio (1 a 14)

'V2005', # Condição no domicílio (responsável, cônjuge, filho, etc.)

'VD2002', # Condição no domicílio (responsável, cônjuge, filho, etc.)

'VD2004', # Espécie da unidade doméstica (1: unipessoal, 2: nuclear, 3: estendida, 4: composta)

**# Membros da família**

'V1014', # Painel

'V2001', # Tamanho da família (nr. pessoas domicílio)

'V2003', # Ordem do morador na família

'V2007', # Sexo do morador

'V20081', # Mês de nascimento do morador

'V20082', # Ano de nascimento do morador

'V2009', # Idade do morador

'V2010', # Cor ou raça do morador

'VD2003', # Nr. de componentes/moradores (1 a 30)

**# Educação**

'V3002', # Frequenta escola?

'V3002A', # Tipo de escola (pública, privada, etc.)

# 'V3003', # Curso atual ou série frequentada (NÃO RODOU 2023)

'V3003A', # Curso atual ou série frequentada

'V3006', # Ano ou série que frequentava anteriormente

# 'V3009', # Maior escolaridade atual do morador (NÃO RODOU 2023)

'V3009A', # Maior escolaridade atual do morador

'VD3006', # Grupos de estudo (1 a 4, 5 a 8 etc.)

**# Trabalho e Rendimento**

'V4001', # Trabalhou na semana de referência?

'V4009', # Quantos trabalhos tinha na semana de referência?

'VD4002', # Se pessoas >= 14 anos estavam ocupadas ou não

'VD4008', # Tipo de vínculo empregatício (empregado, doméstica, militar etc.)

'VD4009', # Tipo de vínculo empregatício (empregado, doméstica, militar etc.)

'VD4010', # Ramo em que trabalha (agricultura, indústria, comércio etc.)

'VD4011', # Grupamentos ocupacionais (diretor, apoio adm., técnico agropecuários etc.)

'VD4013', # Faixa de horas de trabalho semanal

'VD4016', # Rendimento mensal habitual (R$)

'VD4017', # Rendimento mensal efetivo (R$)

'VD4047', # Rendimento efetivo recebido de programas sociais, seguro-desemprego, seguro-defeso, bolsa de estudos, rendimento de caderneta de poupança e outras aplicações financeiras

'V5001A', # Recebe BPC-LOAS?

'V5001A2', # R$ BPC-LOAS

'V5002A', # Recebe Bolsa Família?

'V5002A2', # R$ Bolsa Família

'V5003A', # Outros programas sociais?

'V5003A2', # R$ Outros programas sociais

**# ## Pesos**

'V1031', # Peso domicílio e pessoas SEM calibração

'V1032', # Peso domicílio e pessoas COM calibração

'V1034' # Projeção da população de 1º de julho por sexo e idade

)

**MODELOS ECONOMÉTRICOS – CONSTRUÇÃO CONJUNTA**

**Objetivos do trabalho:**

* **Caracterizar o público-alvo**: Quantificar e entender o perfil socioeconômico dos alunos elegíveis
  + **Método**: Estatística descritiva
* **Analisar padrões de abandono e evasão escolar:** Diferenciar entre alunos que interrompem e aqueles que continuam seus estudos, usando dados de frequência e indicadores de escolaridade ao longo dos trimestres
  + **Método**: Logit / Probit / Heckit
* **Simular o impacto do programa:** Estimar custos
  + **Método**: Logit / Probit / Heckit
* **Estimação da efetividade**: Analisar a efetividade do programa em termos de retenção escolar e estímulo à conclusão do ensino médio
  + **Método**: Logit / Probit / Heckit

**Variáveis dependente**

* Y tal
* Y tal
* ...

**Variáveis independentes**

* x tal
* x tal
* x tal
* x tal
* x tal
* x tal
* ...

Variáveis Curti que estão no meu df

variaveis\_interesse <- c(

"UPA", "V1008", "V1014", "V2003", "V2008", "V20081", "V20082", "V2009",

"V3003A", "VD2004", "V3002A", "ID\_DOMICILIO", "V2001", "VD4016",

"VD4017", "Trimestre", "VD2003", "VD4019", "V2007", "V3009A","VD2002", "V3006", "VD3005", "V2010")

Selecionados Fernando

Inseridos PNADc