PNAD COVID19 – Plano amostral e ponderação

1. OBJETIVOS

A PNAD COVID19 é uma versão da Pesquisa Nacional por Amostra de Domicílios Contínua, realizada em parceria com o Ministério da Saúde, com coleta de dados por telefone. Seus objetivos incluem estimar número de pessoas com sintomas de COVID-19 e quantificar alguns dos impactos da pandemia no mercado de trabalho.

2. PLANO AMOSTRAL

Para a realização da pesquisa PNAD COVID19 foi utilizada como base a amostra de domicílios da Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua do 1º Trimestre de 2019. Para compreender como essa amostra foi utilizada, é importante lembrar como é o plano amostral da PNAD Contínua.

A amostra original da PNAD Contínua[1] foi obtida por um plano amostral conglomerado em dois estágios com estratificação das unidades primárias de amostragem (UPAs). No primeiro estágio foram selecionadas UPAs com probabilidade proporcional ao número de domicílios dentro de cada estrato definido. No segundo estágio foram selecionados 14 domicílios particulares permanentes (que podem estar ou não ocupados) dentro de cada UPA da amostra do primeiro estágio. O sorteio dos domicílios em cada UPA foi feito por amostragem aleatória simples, considerando os endereços listados no Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) atualizado para cada UPA.

Considerando que a PNAD COVID19 deveria ser implementada a partir do mês de maio de 2020, foi escolhida como base a amostra coletada no 1º Trimestre de 2019. O motivo dessa escolha foi que esta amostra inclui somente domicílios que já teriam completado sua participação na PNAD Contínua até o fim do 1º Trimestre de 2020. Além disso, a amostra completa de um trimestre tem um tamanho considerado suficiente para garantir a inferência para os domínios de estimação usualmente considerados na divulgação de resultados da PNAD Contínua, e é razoável esperar que poderia suportar inferência similar para a PNAD COVID19. Essa escolha buscou também evitar sobrecarga nos domicílios que ainda vão responder a PNAD Contínua ao longo do ano de 2020. Uma desvantagem dessa escolha é ter que voltar a

[1] Mais detalhes sobre o planejamento da amostra da PNAD Contínua podem ser encontrados nas Notas metodológicas da Pesquisa Nacional por Amostra de Domicílios Contínua, disponível em https://www.ibge.gov.br/estatisticas/multidominio/condicoes-de-vida-desigualdade-e-pobreza/17270-pnad-continua.html?=&t=downloads

domicílios cuja participação na PNAD Contínua fora dada como encerrada, e com isso aumentar a sobrecarga de coleta sobre esses domicílios.

Com isso, a amostra de domicílios considerada como base tinha 211.000 domicílios. Essa amostra foi submetida a um processo de pareamento para integração com outras bases de dados buscando obter números de telefone para cada domicílio. Ao final desse processo, o número de domicílios com ao menos um telefone disponível para permitir sua inclusão na amostra da PNAD COVID19 ficou em 193.662 domicílios, representando cerca de 92% da amostra base.

O plano para utilização dessa amostra de domicílios considera quatro demandas principais:

a) Produzir estimativas semanais da população segundo ocorrência de sintomas associados à COVID-19 e providências tomadas em caso de sintomas, em níveis nacional e por macrorregião;
b) Produzir estimativas semanais da taxa de desocupação da população de 14 anos e mais, em níveis nacional e por macrorregião;
c) Produzir estimativas mensais da taxa de desocupação e outros indicadores do mercado de trabalho para a população de 14 anos e mais, em níveis nacional e por unidades da federação;
d) Produzir estimativas mensais das variações da taxa de desocupação e outros indicadores do mercado de trabalho para a população de 14 anos e mais, em níveis nacional e por unidades da federação.

Para dar conta das demandas a) e b), a amostra total da pesquisa foi dividida em 4 semanas, de maneira a obter resultados semanais. Para dar conta da demanda d), o planejamento da pesquisa é do tipo painel, com a amostra da pesquisa sendo mantida fixa ao longo dos meses que a pesquisa for repetida. A princípio, pretende-se repetir a pesquisa por quatro meses, podendo essa duração ser estendida dependendo da duração da pandemia.

Para a alocação dos domicílios da amostra por semana de coleta foi empregado um método de amostragem estratificada simples de domicílios, com alocação igual. Esse método foi implementado mediante um algoritmo em três passos. No primeiro passo, foi atribuído a cada domicílio da amostra base, de forma independente, um número aleatório permanente gerado segundo a distribuição Uniforme (0,1). No segundo passo, dentro de cada estrato de seleção da amostra base, os domicílios foram então ordenados em forma crescente segundo os números aleatórios gerados no passo 1. No terceiro passo, dentro de cada estrato de seleção da

amostra base, os domicílios assim ordenados foram alocados em quatro subgrupos de tamanhos iguais, ou aproximadamente iguais, quando o número de domicílios na amostra base do estrato não era múltiplo de quatro. Os quatro subgrupos de domicílios assim definidos formaram as amostras a serem coletadas em cada uma das quatro semanas do período de coleta.

Para fins da inferência mensal, a amostra da PNAD COVID19 seria idêntica à amostra de um trimestre típico da PNAD Contínua, a menos da não resposta causada:

a) Pela não obtenção de um número de telefone para uma parte (8%) dos domicílios da amostra base;
b) Pela perda de domicílios durante a coleta por telefone da amostra com telefones disponíveis.

Já para a inferência semanal, a amostra da PNAD COVID19 para cada uma das quatro semanas pode ser pensada como obtida através de um plano com amostragem em duas fases:

a) na primeira fase, a amostra de domicílios decorre do sorteio para a amostra da PNAD Contínua;
b) na segunda fase, para cada semana ocorre o sorteio de domicílios da amostra inicial por meio de amostragem estratificada simples de domicílios, com alocação igual, sendo usados os mesmos os estratos de seleção da amostra da PNAD Contínua.

Isto implica que há necessidade de considerar métodos adicionais para ponderação e estimação de variâncias no caso das estimativas semanais, como será descrito na sequência. Note-se que também as amostras semanais estarão sujeitas às perdas por não resposta já descritas, e portanto, que será também necessário compensar estas perdas através da ponderação a ser empregada para obtenção das estimativas.

3. PONDERAÇÃO DA AMOSTRA

O processo de ponderação da amostra terá quatro ou cinco etapas, conforme o período de referência da amostra das estimativas. No caso de estimativas mensais (ou obtidas mediante acumulação das amostras de quatro semanas consecutivas), esse processo terá somente quatro etapas. Para as estimativas com base em amostras de apenas uma semana, será necessária uma etapa adicional para incorporar o efeito do sorteio da amostra da segunda fase, como já indicado. As etapas do processo de ponderação proposto incluem:

1. Calcular peso básico de seleção do domicílio para a PNAD Contínua;
2. Calcular fator de ajuste do peso básico para não resposta devida à perda de domicílios no pareamento para obtenção de telefones de contato;

3. Calcular fator de ajuste do peso básico devido ao sorteio para a segunda fase (amostra semanal);
4. Calcular fator de ajuste para não resposta devida à perda durante a coleta de dados da PNAD COVID19;
5. Calibração do peso para totais populacionais conhecidos.

Na sequência, as etapas do processo de ponderação são descritas com mais detalhes.

Cálculo do peso básico de seleção do domicílio

Considerando a descrição do plano amostral da PNAD Contínua conforme IBGE (2020), o peso amostral básico de um domicílio é dado por:

$$d_{hij} = (1 / \pi_{hi}) \times (1 / \pi_{j|hi}) \quad (1)$$

onde d_{hij} é o peso básico do domicílio *j* da UPA *i* do estrato *h* da amostra da PNAD Contínua no primeiro trimestre de 2019, π_{hi} é a probabilidade de inclusão na amostra da UPA *i* do estrato *h* da amostra da PNAD Contínua no primeiro trimestre de 2019, e $\pi_{j|hi}$ é a probabilidade condicional de inclusão do domicílio *j* na amostra da UPA *i* do estrato *h* da amostra da PNAD Contínua no primeiro trimestre de 2019.

Cálculo do primeiro fator de ajuste de não resposta

O primeiro fator de ajuste de não resposta visa compensar a perda de domicílios da amostra inicial causada pelas perdas do processo de pareamento para obtenção de ao menos um número de telefone para viabilizar a pesquisa telefônica. Na ausência de revisão do processo de pareamento para obter números de telefone, este fator de ajuste poderia ser calculado uma única vez, e considerando a amostra inteira.

O método usado inicialmente para obter este fator de ajuste foi um 'ajuste de não resposta por classes', onde as classes foram os estratos de seleção da amostra da PNAD Contínua. Este método está descrito, por exemplo, em (SÄRNDAL; SWENSSON; WRETMAN, 2003, p. 580), e corresponde a usar pesos ajustados w_{hij} definidos como:

$$w_{hij} = d_{hij} \times (1 / T_h) \quad (2)$$

onde T_h é a proporção de domicílios da amostra base no estrato *h* com telefone cadastrado.

Cálculo do fator de ajuste do peso básico devido ao sorteio para a segunda fase (amostra semanal)

Como o sorteio para a amostra de cada semana foi feito por amostragem estratificada simples dentro de cada estrato de seleção da amostra base, e como para cada semana foram

alocados aproximadamente um de cada quatro domicílios da amostra base, este fator de ajuste é essencialmente igual a quatro. Porém, como os tamanhos de amostra de domicílios ao longo dos vários estratos de seleção da amostra base não são sempre múltiplos de quatro, podem ocorrer pequenas variações neste fator de ajuste. Então, o peso ajustado para domicílios na estimação de amostras semanais é dado por:

$$w_{hij}^{s} = w_{hij} \times (n_h / n_h^s) \quad (3)$$

onde n_h^s é o número de domicílios com telefone no estrato *h* alocados na amostra da semana *s*, e $n_h = \sum_{s=1}^{4} n_h^s$ é o número total de domicílios com telefone na amostra do estrato *h*.

Cálculo do fator de ajuste para não resposta devida à perda durante a coleta de dados

Este cálculo terá que ser feito de maneira distinta, dependendo de qual será a amostra cuja não resposta precisa ser compensada. Para o caso da amostra mensal, o peso ajustado para não resposta devida à perda durante a coleta de dados deve ser obtido por:

$$w_{hij}^{*} = w_{hij} \times (n_h / m_h) \quad (4)$$

onde m_h é o número de domicílios com telefone e com entrevista realizada na amostra do estrato *h*.

Para o caso de amostra semanal, o peso ajustado para não resposta devida à perda durante a coleta de dados deve ser obtido por:

$$w_{hij}^{s,a} = w_{hij}^{s} \times (n_h^s / m_h^s) \quad (5)$$

onde m_h^s é o número de domicílios com telefone e com entrevista realizada na amostra da semana *s* do estrato *h*.

Calibração do peso para totais populacionais 'conhecidos'

A última etapa de cálculo dos pesos é a calibração para totais populacionais 'conhecidos', feita mediante pós-estratificação simples. Os pós-estratos considerados nessa etapa correspondem a grupos de sexo combinados às faixas de idade de forma a recompor a estrutura demográfica da população por Unidade da Federação. Os totais populacionais utilizados são as estimativas populacionais fornecidas pela Coordenação de População e Indicadores Sociais[2], calculadas para cada Unidade da Federação, por sexo e faixas etárias

[2] Populações para o dia 1º de cada mês disponíveis em: ftp://ftp.ibge.gov.br/Projecao_da_Populacao/Projecao_da_Populacao_2018/Populacoes_Projetadas_Mensais_dia_01_ate_2030.xlsx

decenais com referência no sábado da semana imediatamente anterior, no caso das estimativas semanais, e com referência no dia 15 do mês de coleta, para o caso das estimativas mensais.

O peso final para cada domicílio com entrevista realizada na amostra mensal é calculado da seguinte forma:

$$w_{hij}^{c} = w_{hij}^{*} \times (P_g^m \,/\, \hat{P}_g^m) \qquad (6)$$

onde P_g^m é o total populacional do pós-estrato g para o dia 15 do mês m, e $\hat{P}_g^m$ é a estimativa desse total populacional obtida com os dados da pesquisa usando os pesos w_{hij}^{*}.

O peso final para cada domicílio com entrevista realizada na amostra semanal é calculado da seguinte forma:

$$w_{hij}^{s,c} = w_{hij}^{s,a} \times (P_g^s \,/\, \hat{P}_g^s) \qquad (7)$$

onde P_g^s é o total populacional do pós-estrato g para a semana s, e $\hat{P}_g^s$ é a estimativa desse total populacional obtida com os dados da pesquisa usando os pesos $w_{hij}^{s,a}$.

4. APRIMORAMENTO DA CORREÇÃO DA NÃO RESPOSTA

Os ajustes para não resposta descritos nas expressões (2), (4) e (5) representam correções bastante simples, mas que dependem de um modelo para a propensão a responder à pesquisa que estabelece que essa propensão varia apenas com os estratos de seleção da amostra. Esse modelo não foi analisado e validado, nesta primeira ocasião, por razões de tempo. Uma possibilidade que deve ser explorada é enriquecer este modelo considerando variáveis medidas em nível de domicílio disponíveis da coleta dessa amostra na PNAD Contínua.

Para poder avaliar a possível melhoria do ajuste dos pesos, é necessário contar com acesso aos microdados da PNAD Contínua para todos os domicílios da amostra do 1º trimestre de 2019, acrescentados de indicadores de:

a) Existência de ao menos um telefone cadastrado para cada domicílio;
b) Entrevista realizada ou não na coleta da pesquisa;
c) Semana de coleta em que o domicílio foi alocado.

De posse dessas informações, será possível ajustar modelos de regressão logística para a propensão a responder. Considerando as propensões estimadas por tais modelos, será então possível calcular fatores de correção da não resposta com melhor desempenho para redução dos possíveis vieses devidos à não resposta diferencial potencialmente observada na pesquisa.

5. ESTIMAÇÃO DE VARIÂNCIAS

As estimativas de variância para as quantidades populacionais de interesse na PNAD COVID19 são obtidas mensalmente através do método do Conglomerado Primário (Cochran 1977, p.307). Além disso, como estas quantidades são obtidas através de estimadores do tipo razão, não existe uma expressão exata para suas respectivas variâncias. Sendo assim, é utilizada a técnica de Linearização de Taylor para obter uma variância aproximada.

No caso das estimativas semanais utiliza-se uma aproximação do método do Conglomerado Primário baseando-se nas UPAs associadas aos domicílios selecionados na segunda fase da pesquisa.

6. ESTIMAÇÃO EM NÍVEL DOMICILIAR

Inicialmente, a estimação de variáveis em nível domiciliar na PNAD COVID19 é feita por domínio de interesse (Cochran 1977, p.36). Neste caso, uma variável y que pertencente ao domínio passa a ser representada da seguinte forma.

$$y_{hij}^{d} = \begin{cases} y_{hij}^{d} & i \in D \\ 0 & i \notin D \end{cases} \quad (8)$$

Onde o domínio de interesse *D* é definido pelo responsável pelo domicílio. Neste caso, as características domiciliares são ponderadas pelo fator de expansão do responsável pelo domicílio.

Ao calibrar os pesos amostrais de forma a recompor a estrutura demográfica da população e utilizar um estimador do tipo razão, os mesmos passam a não ser constantes dentro do mesmo domicílio. Neste sentido, estudos estão sendo realizados de forma a avaliar o impacto de utilizar este tipo de abordagem metodológica para estimação em nível domiciliar. Neste sentido, novos estimadores podem ser propostos de forma a melhorar o processo de estimação pontual e de cálculo dos erros amostrais associados as estimativas da pesquisa.

7. REFERÊNCIAS

COCHRAN, W.G. Sampling Techniques. 3rd. ed. New York: John Wiley, 1977.

IBGE (2020). Pesquisa Nacional por Amostra de Domicílios Contínua - Notas técnicas - Versão 1.7. Rio de Janeiro: IBGE.

SÄRNDAL, Carl-Erik; SWENSSON, Bengt; WRETMAN, Jan. **Model assisted survey sampling**. New York: Springer, 1992.