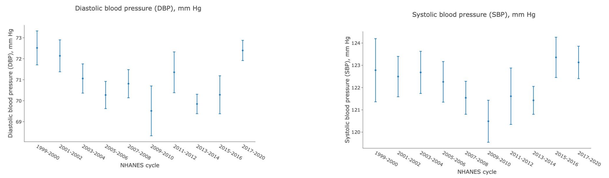
COLUMBIA
MAILMAN SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS

Possible Correlates of Decreasing Trend of BP Control Among U.S. Adults with Hypertension from 1999 to 2020: A Bayesian Approach

Lincole Jiang & Yanran Li

OBJECTIVE

Controlling blood pressure (BP) reduces the risk for cardiovascular disease. However, the prevalence of BP control (i.e., systolic BP < 140 and diastolic BP < 90) among US adults with hypertension has decreased since 2013. The intention of this study is to analyze publicly available data from US adults to help identify potential causes or correlates of worsening BP control among US adults with hypertension over the past decade, as this may allow for development of effective interventions to help control BP and prevent cardiovascular disease.



BACKGRUOND

Hypertension, commonly known as high blood pressure, is a critical public health concern globally, particularly in the United States. It is a major risk factor for cardiovascular diseases (CVDs), the leading cause of death in the U.S. Effective blood pressure (BP) control, defined as maintaining systolic BP below 140 mmHg and diastolic BP below 90 mmHg[1], is essential for reducing the risk of these diseases. Despite the known benefits of BP control, recent trends indicate a worrying decline in the prevalence of controlled BP among U.S. adults with hypertension.

METHODS

- **Model Specification**

We considered a hierarchical Bayesian GLM for this exercise. Explicitly, denoting p as the number of input features (p=22), n as the number of groups of data (n=10), m_i as the number of datapoints in each group, λ as the hyperprior, $\pi(y)$, $t(y)$, $a_y(\eta)$ as the exponential family for the response, and η_{μ} as the mean mapping for the response family, f as the canonical link function, and $x_{ij} \in \mathrm{R}^p$ as feature vectors, we generate the model as:

1. Draw the prior parameter
$$\lambda \sim \text{gamma}(\eta, 1)$$
2. For each group i:
   (a) Draw coefficients
   $$\beta_i \mid \lambda \sim \mathrm{N}(0, \lambda^2)$$
   (b) For each datapoint j, draw
   $$y_{ij} \mid x_{ij}, \beta_i \sim \text{expfam}(\eta_{ij}) \qquad \eta_{ij} = \eta_\mu(f(\beta_i \cdot x_{ij}))$$

In particular, the exponential family we considered here is the Bernoulli-Beta exponential family, and denoting θ as the probability of outcome, i.e., having hypertension, we have:

$$\eta = \log(\frac{\theta}{1-\theta}), t(x) = x, \pi(x) = 1, a(\eta) = -\log(1-\theta) = \log(1+e^{\eta})$$

METHODS

- **Algorithm**

We implemented both ADVI and HMC for this model, using built-in features in `stan`. The algorithm for ADVI is specified in Algorithm 1 of the Kucukelbir 2017 paper, whereas the No-U-Turn sampler, which was the default MCMC algorithm implemented in `stan` and a variant of HMC, was detailed as Algorithm 2 in the 2014 Hoffman and Gelman paper.

Algorithm 1: Automatic differentiation variational inference (ADVI)

Input: Dataset $\mathbf{x} = x_{1:N}$, model $p(\mathbf{x}, \boldsymbol{\theta})$.
Set iteration counter $i = 1$.
Initialize $\boldsymbol{\mu}^{(1)} = \mathbf{0}$.
Initialize $\boldsymbol{\omega}^{(1)} = \mathbf{0}$ (mean-field) or $\mathbf{L}^{(1)} = \mathbf{I}$ (full-rank).
Determine η via a search over finite values.
while *change in* ELBO *is above some threshold* **do**
  Draw M samples $\boldsymbol{\eta}_m \sim \text{Normal}(\mathbf{0}, \mathbf{I})$ from the standard multivariate Gaussian.
  Approximate $\nabla_{\boldsymbol{\mu}} \mathcal{L}$ using MC integration (Equation (7)).
  Approximate $\nabla_{\boldsymbol{\omega}} \mathcal{L}$ or $\nabla_{\mathbf{L}} \mathcal{L}$ using MC integration (Equations (8) and (9)).
  Calculate step-size $\boldsymbol{\rho}^{(i)}$ (Equation (10)).
  Update $\boldsymbol{\mu}^{(i+1)} \longleftarrow \boldsymbol{\mu}^{(i)} + \text{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\boldsymbol{\mu}} \mathcal{L}$.
  Update $\boldsymbol{\omega}^{(i+1)} \longleftarrow \boldsymbol{\omega}^{(i)} + \text{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\boldsymbol{\omega}} \mathcal{L}$ or $\mathbf{L}^{(i+1)} \longleftarrow \mathbf{L}^{(i)} + \text{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\mathbf{L}} \mathcal{L}$.
  Increment iteration counter.
end
Return $\boldsymbol{\mu}^* \longleftarrow \boldsymbol{\mu}^{(i)}$.
Return $\boldsymbol{\omega}^* \longleftarrow \boldsymbol{\omega}^{(i)}$ or $\mathbf{L}^* \longleftarrow \mathbf{L}^{(i)}$.

Algorithm 2 Naive No-U-Turn Sampler

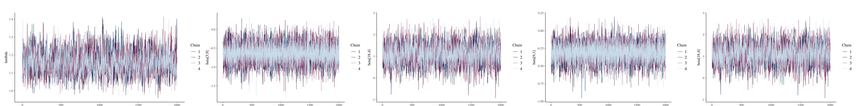
Given θ^0, ϵ, $\mathcal{L}$, M:
for $m = 1$ to M **do**
  Resample $r^0 \sim \mathcal{N}(0, I)$.
  Resample $u \sim \text{Uniform}([0, \exp\{\mathcal{L}(\theta^{m-1}) - \frac{1}{2} r^0 \cdot r^0\}])$
  Initialize $\theta^- = \theta^{m-1}$, $\theta^+ = \theta^{m-1}$, $r^- = r^0$, $r^+ = r^0$, $j = 0$, $\mathcal{C} = \{(\theta^{m-1}, r^0)\}$, $s = 1$.
  while $s = 1$ **do**
    Choose a direction $v_j \sim \text{Uniform}(\{-1, 1\})$.
    if $v_j = -1$ **then**
      $\theta^-, r^-, \_, \_, \mathcal{C}', s' \leftarrow \text{BuildTree}(\theta^-, r^-, u, v_j, j, \epsilon)$.
    else
      $\_, \_, \theta^+, r^+, \mathcal{C}', s' \leftarrow \text{BuildTree}(\theta^+, r^+, u, v_j, j, \epsilon)$.
    end if
    if $s' = 1$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$.
    end if
    $s \leftarrow s' \mathbb{I}[(\theta^+ - \theta^-) \cdot r^- \geq 0] \mathbb{I}[(\theta^+ - \theta^-) \cdot r^+ \geq 0]$.
    $j \leftarrow j + 1$.
  end while
  Sample θ^m, r uniformly at random from $\mathcal{C}$.
end for

function BuildTree($\theta, r, u, v, j, \epsilon$)
if $j = 0$ **then**
  Base case—take one leapfrog step in the direction v.
  $\theta', r' \leftarrow \text{Leapfrog}(\theta, r, v\epsilon)$.
  $\mathcal{C}' \leftarrow \begin{cases} \{(\theta', r')\} & \text{if } u \leq \exp\{\mathcal{L}(\theta') - \frac{1}{2} r' \cdot r'\} \\ \emptyset & \text{else} \end{cases}$
  $s' \leftarrow \mathbb{I}[\mathcal{L}(\theta') - \frac{1}{2} r' \cdot r' > \log u - \Delta_{\max}]$.
  return $\theta', r', \theta', r', \mathcal{C}', s'$.
else
  Recursion—build the left and right subtrees.
  $\theta^-, r^-, \theta^+, r^+, \mathcal{C}', s' \leftarrow \text{BuildTree}(\theta, r, u, v, j-1, \epsilon)$.
  if $v = -1$ **then**
    $\theta^-, r^-, \_, \_, \mathcal{C}'', s'' \leftarrow \text{BuildTree}(\theta^-, r^-, u, v, j-1, \epsilon)$.
  else
    $\_, \_, \theta^+, r^+, \mathcal{C}'', s'' \leftarrow \text{BuildTree}(\theta^+, r^+, u, v, j-1, \epsilon)$.
  end if
  $s' \leftarrow s' s'' \mathbb{I}[(\theta^+ - \theta^-) \cdot r^- \geq 0] \mathbb{I}[(\theta^+ - \theta^-) \cdot r^+ \geq 0]$.
  $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{C}''$.
  return $\theta^-, r^-, \theta^+, r^+, \mathcal{C}', s'$.
end if

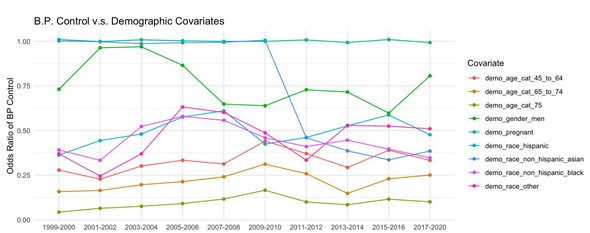
RESULTS/OUTCOMES

- **Convergence Diagnostics**

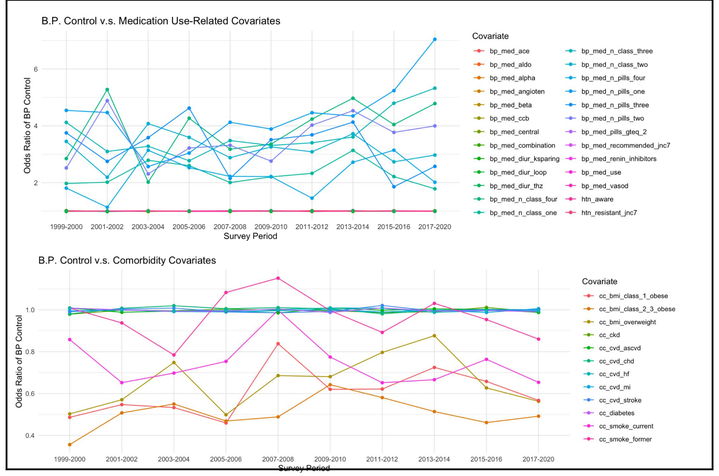
For ADVI, convergence was poor with Pareto K-value ranging from 3.6 and 4.4 despite adjustments in number of iterations and convergence tolerance. For MCMC, Gelman-Rubin statistics ranges from within plut or minus one thousandths of 1, indicating very good convergence. The traceplots for a few selected parameters show the same:



- **Interpretation of Coefficients**



RESULTS/OUTCOMES



DATA / DISCUSSION

The dataset considered is directly accessible through R-package “cardioStatsUSA”, for which contains processed hypertension-related data from National Health and Nutrition Examination Survey (NHANES) of ten survey periods from 1999 to March 2020, including hypertension status, demographic variables like age, height, race; BP-related medication use; as well as comorbidity variables including . Most variables from this R-package was kept and only repeated variables were excluded. Given the hypothesis, only the observations with at least one measure of blood pressure and the subpopulation with hypertension (defined in JNC7-guideline) was kept. Our approach follows the Bayesian probabilistic modeling framework as an iterative process illustrated by Boxer’s loop. As of now, we are still in the process of repeating the process of model building, experimenting with different parameters, and trying different classes of models to provide a superior solution. This will be our main objective for the remainder of this semester (we do not that the nature of the dataset makes sequential point models inapplicable). After making a selection of the most significant correlates and given the nature of our hypothesis, we attempt also to employ causal inference methods like regression discontinuity design (RDD) to identify the causal correlates of decreasing BP control after 2013, albeit we recognize that the nature of the working data as cross-sectional makes it difficult to establish exchangeability in causal models.

REFERENCES

Byron Casey Jaeger (2022). NHANES Data, 1999 - 2020}
https://jhs-hwg.github.io/cardioStatsUSA/reference/nhanes_data.html
Hoffman, Matthew D., and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." J. Mach. Learn. Res. 15.1 (2014): 1593-1623.
Kucukelbir, Alp, et al. "Automatic differentiation variational inference." Journal of machine learning research (2017).
Blei. Lecture Notes. (2023)
Blei. “Build, compute, critique, repeat: Data analysis with latent variable models.” Annual Review of Statistics and Its Applications 1 (2014): 203-232