HMS520-Autumn 2025 Final Project

Modelling Systolic Blood Pressure by Body Mass Index in National Health and Nutrition Examination Survey (2021-2023)

Zaw Wai Yan Bo, Pyone Yadanar Paing

BE BOUNDLESS

Format and overall goals of the project

- To analyze National Health and Nutrition Examination Survey (NHANES) dataset that uses data wrangling and modeling tools in R.
- To model mean systolic blood pressure SBP (mmHg) by body mass index B(kg/m^2) among adults 20+ years old in the United States by using NHANES (2021-2023) dataset.
- BMI is considered a key determinant of SBP, and its importance differs by age group.

GitHub Repository Overview

Structure:

- Script file: hms520_final_project_code.R
- Datasets folder
- Figures folder
- Presentation slide PDF
- README with project documentation

Goals:

- Completeness
- Reproducibility
- Code readability and extendibility

Description of the dataset

- NHANES provides cross-sectional data of the health and nutrition of the United States population.
- Uses a complex, multistage probability sampling design to produce nationally representative estimates
- Household interviews
- Physical examinations
- Laboratory tests

Load libraries

- Used "haven" package for loading xpt files.
  - library(haven)
- Used "data.table" package for data wrangling.
  - library(data.table)
- Used "dplyr" package for selecting, filtering and mutating.
  - library(dplyr)
- Used "ggplot2" package for plotting
  - library(ggplot2)
- Used "rlang" for symbolize and inject of variable names in a function
  - library(rlang)

Download and load datasets

- Download and load three datasets from NHANES website.
- BMI dataset
  - bmi <- read_xpt("BMX_L.xpt")
- BP dataset
  - bp <- read_xpt("BPXO_L.xpt")
- Demographic dataset
  - demo <- read_xpt("DEMO_L.xpt")
- Merge those three datasets by respondent ID.
  - nhanes <- merge(merge(bp, bmi, by = "SEQN", all = TRUE),
    demo, by = "SEQN", all = TRUE)

Data cleaning

- Transformed the dataset into a data table.
  - nhanes_dt <- setDT(nhanes)
- Subset the data for individuals 20 years and older.
  - nhanes_dt <- nhanes_dt[RIDAGEYR >= 20]
- Computed mean SBP by using available
  - nhanes_dt[, mean_sbp := rowMeans(.SD, na.rm = TRUE), .SDcols = c("BPXOSY1", "BPXOSY2", "BPXOSY3")]
- Created age groups
  - nhanes_dt[, age_gp := cut(RIDAGEYR, breaks = c(20, 35, 50, 65, Inf), labels = c("20–34", "35–49", "50-64", "65+"), right = FALSE)]

Data cleaning

- Create BMI categories
  - nhanes_dt[, bmi_cat := ifelse(BMXBMI < 18.5, "Underweight",
    ifelse(BMXBMI < 25, "Healthy weight",
    ifelse(BMXBMI < 30, "Overweight",
    "Obesity")))]
- Limit the dataset for non-missing rows of age, BMI and BP
  - nhanes_final <- nhanes_dt[!is.na(mean_sbp) & !is.na(BMXBMI) & !is.na(RIDAGEYR]

Plotting and Summary Statistics

> Function to create histogram (SBP by age, BMI by age, SBP by BMI)

```
hist_plot <- function(data, xvar, facet_var, binwidth = 1,
                      fill = "skyblue", color = "black") {

  ggplot(data, aes_string(x = xvar)) +
    geom_histogram(binwidth = binwidth, fill = fill, color = color) +
    facet_wrap(as.formula(paste("~", facet_var)))
}
```

Create scatterplot (SBP by BMI, SBP by BMI across age groups)

```
plot_xy <- function(data, xvar, yvar, facet_var = NULL) {
  # Convert character variable names to symbols
  x <- sym(xvar)
  y <- sym(yvar)

  p <- ggplot(data, aes(x = !!x, y = !!y)) +
    geom_point(alpha = 0.6) +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(
      title = paste(yvar, "vs", xvar),
      x = xvar, y = yvar
    ) # Optional facet
if (!is.null(facet_var)) {
    facet_sym <- sym(facet_var)
    p <- p + facet_wrap(vars(!!facet_sym))
  } return(p)}
```

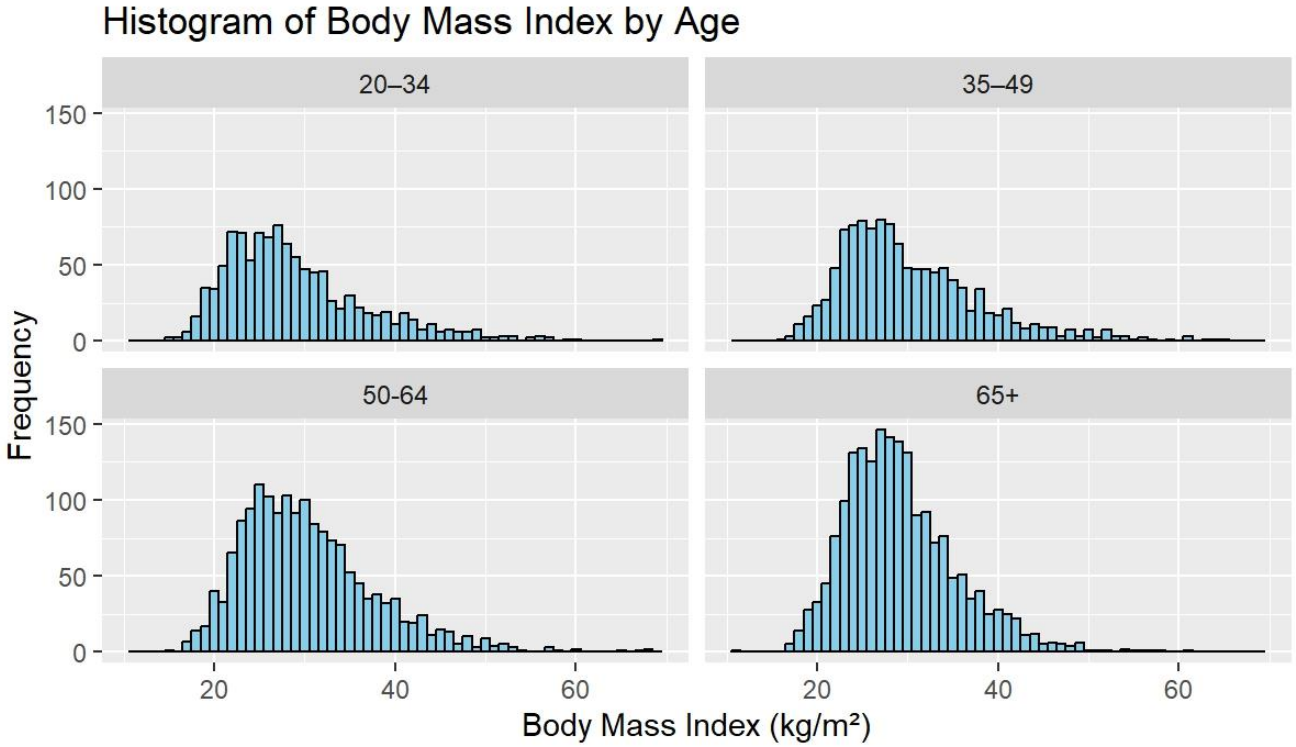
Modeling Approach

- Fit Linear Regression Model of SBP on BMI.
  - Simple Linear Regression Model
    - lm_model <- lm(mean_sbp~BMXBMI, data = nhanes_final)
  - Adjusted by age
    - lm_model_adjusted <- lm(mean_sbp~BMXBMI+RIDAGEYR, data = nhanes_final)

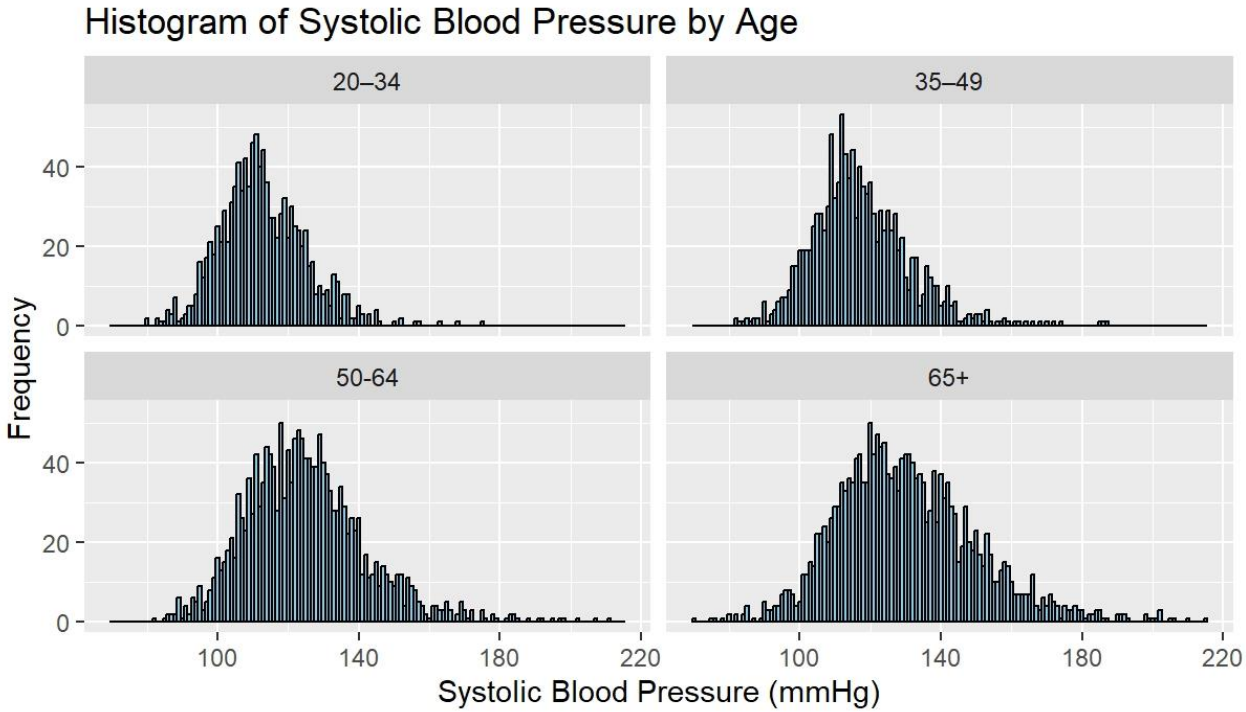
Modeling Approach

- Fit Group-Specific Regression Model of SBP on BMI by Age-groups
  - Split the dataset by age groups
    - nhanes_group <- split(nhanes_final, nhanes_final$age_gp)
  - Fit model for SBP on BMI by age groups
    - models <- lapply(nhanes_group, function(dt) lm(mean_sbp ~ BMXBMI, data = dt))
    - for (key in names(nhanes_group)) { nhanes_group[[key]][, mean_sbp_fit := predict(models[[key]], nhanes_group[[key]])] }
  - Recombine into main dataset
    - nhanes_final <- rbindlist(nhanes_group)

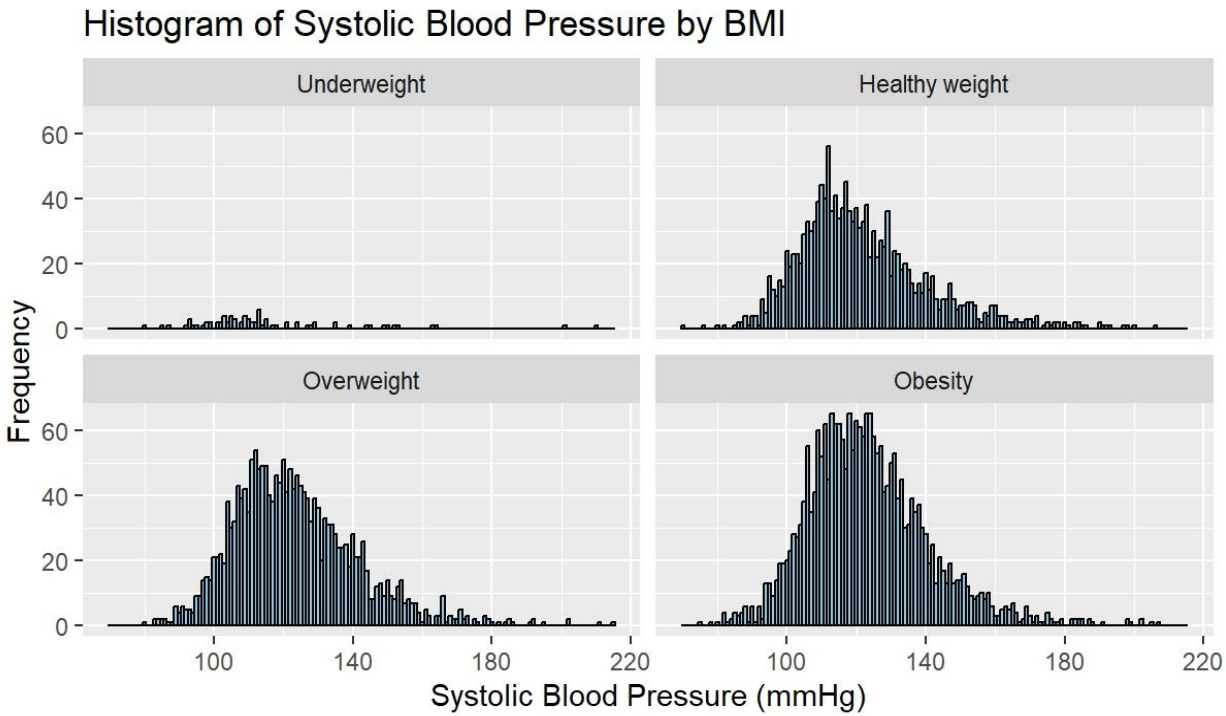
Results: Summary Statistics



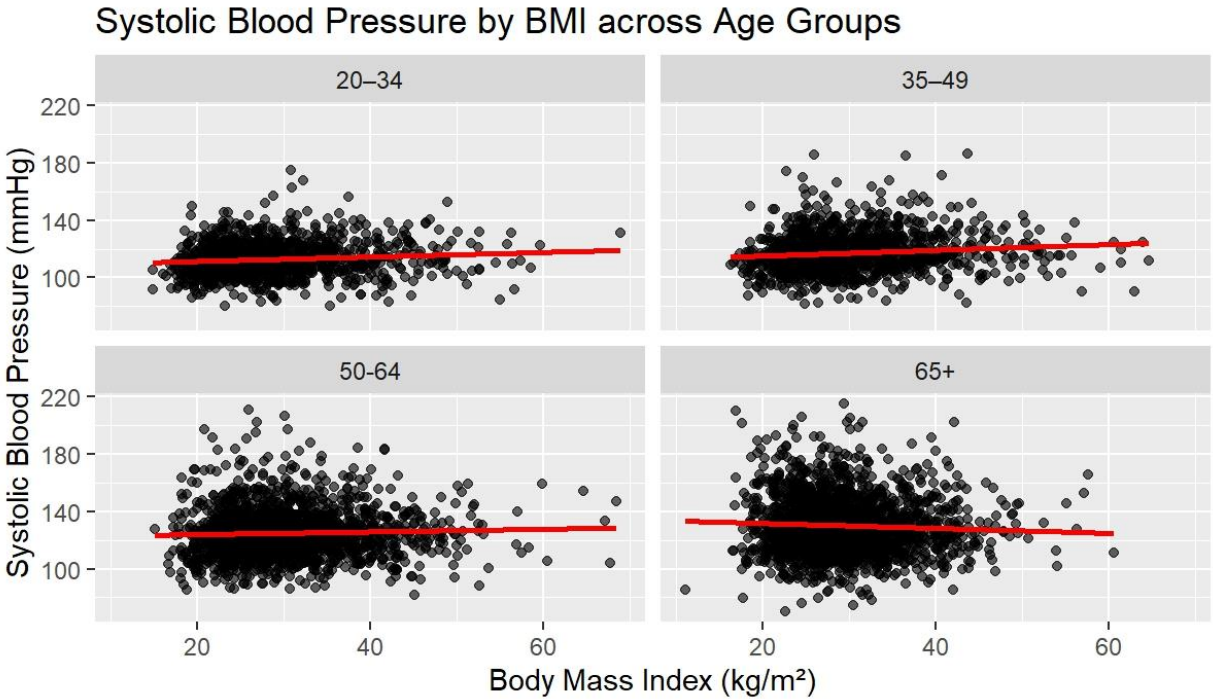
Results: Summary Statistics



Results: Summary Statistics

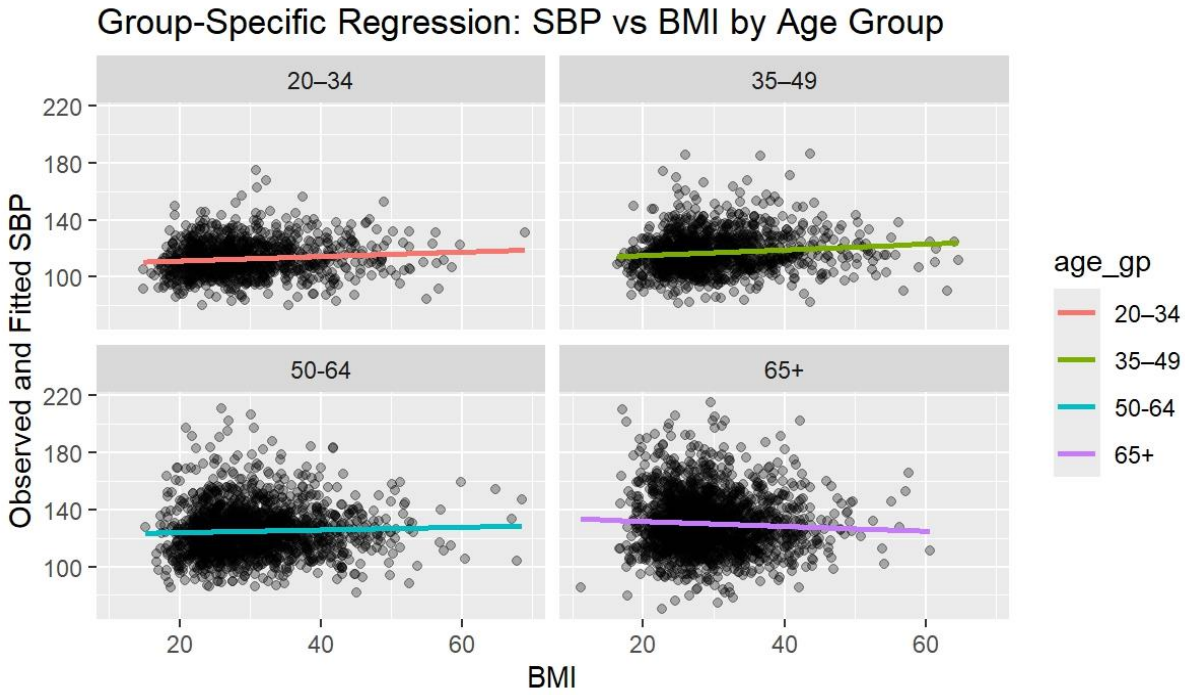


Results: Plots



of WASHINGTON

Results: Model



Results:
Simple Linear Regression Model: SBP on BMI

- Unadjusted

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	120.82291	1.00961	119.673	<2e-16	***
BMXBMI	0.06883	0.03302	2.085	0.0371	*

- Age adjusted

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	98.99372	1.15723	85.544	<2e-16	***
BMXBMI	0.06876	0.03047	2.257	0.0241	*
RIDAGEYR	0.40538	0.01275	31.806	<2e-16	***

Results:
Group-Specific Regression Model: SBP on BMI by Age Groups

- `20–34`
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	108.19036	1.41454	76.484	< 2e-16	***
BMXBMI	0.15824	0.04715	3.356	0.000819	***
- `35–49`
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	111.18173	1.63884	67.842	< 2e-16	***
BMXBMI	0.19878	0.05251	3.785	0.000161	***
- `50-64`
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	122.14990	1.80496	67.675	<2e-16	***
BMXBMI	0.09338	0.05777	1.616	0.106	
- `65+`
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	135.2331	2.2026	61.397	<2e-16	***
BMXBMI	-0.1718	0.0736	-2.334	0.0197	*

Thank you! Questions?

UNIVERSITY *of* WASHINGTON