NHANES Hypertension Analysis (2017)

Author: BharathReddy
Date: 2024-05-05

---

1. Objective

The aim of this study is to analyze the relationship between hypertension and various demographic and biometric factors such as Body Mass Index (BMI), age, income-to-poverty ratio, and gender using data from the NHANES 2017–2018 cycle. Additionally, we build a logistic regression model to predict hypertension status and assess its performance using key classification metrics.

---

2. Dataset Description

Source: National Health and Nutrition Examination Survey (NHANES) 2017–2018, Centers for Disease Control and Prevention (CDC)

Datasets Used:

- DEMO_J.XPT: Demographics (e.g., age, gender, income)
- BMX_J.XPT: Body measurements (e.g., BMI)
- BPX_J.XPT: Blood pressure measurements (4 systolic readings)
- MCQ_J.XPT: Self-reported medical conditions (e.g., diagnosed hypertension)

Final Dataset Size After Cleaning: 4,313 complete observations (with no missing values for relevant variables)

---

3. Target Variable: Hypertension

We defined hypertension using the following criteria:

- **Self-reported diagnosis:** MCQ160B = 1
- **OR** average systolic blood pressure (mean of BPXSY1–BPXSY4) ≥ 130 mmHg

A binary variable hypertension was created, with:

- 1 = hypertensive
- 0 = non-hypertensive

---

4. Exploratory Data Analysis (EDA)

4.1 BMI vs. Hypertension

A boxplot of BMI by hypertension status revealed that individuals with hypertension tend to have higher BMI values. This was statistically confirmed by a Welch Two Sample **t-test**:

- **p-value:** < 0.0001
- **Conclusion:** The mean BMI of hypertensive individuals is significantly higher than non-hypertensive individuals.

4.2 Gender vs. Hypertension

A bar chart showing proportions by gender indicated that a higher percentage of males are hypertensive. A **Chi-square test** of independence was conducted:

- **p-value:** 0.005
- **Conclusion:** There is a statistically significant association between gender and hypertension.

---

5. Predictive Modeling: Logistic Regression

We trained a logistic regression model using the following predictors:

- **BMXBMI**: Body Mass Index
- **RIDAGEYR**: Age
- **INDFMPIR**: Income-to-poverty ratio
- **RIAGENDR**: Gender (1 = male, 2 = female)

Model Summary:

Predictor	Odds Ratio	Interpretation
BMI	1.05	Each unit increase in BMI increases odds by 5%
Age	1.06	Older age significantly raises hypertension risk
Income	0.90	Higher income slightly lowers hypertension risk
Gender (M)	0.82	Males have slightly lower odds when adjusted

All coefficients were statistically significant ($p < 0.01$).

---

6. Model Performance Evaluation

Threshold Used: 0.5 and 0.3 (for sensitivity comparison)

Confusion Matrix at 0.5:

	Predicted No	Predicted Yes
Actual No	2,117	529
Actual Yes	726	942

Metrics:

- **Accuracy:** 70.9%
- **Precision:** 64.0%
- **Recall:** 56.5%
- **F1 Score:** 60.0%

After lowering threshold to 0.3:

- **Accuracy:** 68.2%
- **Recall increased to:** 83.9%
- **F1 Score:** 67.1%

This indicates better sensitivity for detecting hypertensive individuals.

ROC Curve Analysis:

- The Area Under the Curve (AUC) = **0.77**, suggesting strong discriminatory ability of the model.

---

7. Conclusion

This analysis demonstrates that hypertension in adults is strongly associated with increasing age and BMI, with modest influence from gender and income. The logistic regression model, though simple, achieved a good balance between accuracy and sensitivity, with an AUC of 0.77.

Such findings reinforce public health messages around weight control and age-based screening. The model can serve as a useful screening tool for early identification of individuals at risk of hypertension using easily measurable parameters.