ECON 5336 - Project Report
Modeling the Impact of Socioeconomic and Health Factors on Systolic Blood Pressure in Young Adults

Yamini Suguna Kota

1 Introduction

Systolic blood pressure (SBP) is a key indicator of cardiovascular health and a major risk factor for heart disease and stroke. A growing body of literature shows that the early origins of elevated SBP and hypertension can be traced back to adolescence, with factors such as body mass index (BMI), socioeconomic status, and adiposity playing influential roles (Brummett et al. [2019], Ford et al. [2008], Brummett et al. [2012]). Notably, elevated adolescent BMI has been strongly and independently associated with higher SBP and cholesterol in young adulthood (Li et al. [2020]), while socioeconomic status disparities, particularly financial instability, further amplify hypertension risk, especially among racial and ethnic minorities (Colhoun et al. [1998]). Racial and gender-based variations in the relationship between adiposity and SBP also suggest complex interactions that warrant targeted investigation (Zamora-Kapoor et al. [2021]). Moreover, longitudinal analyses reveal that socioeconomic status influences SBP outcomes through both direct and indirect mechanisms, including obesity and lifestyle behaviors, with the association being particularly strong in women and low-income groups (Brummett et al. [2011], Brummett et al. [2012]).

Motivated by this literature, the objective of this study is to investigate the impact of several lifestyle and demographic variables on systolic blood pressure (SBP) using Ordinary Least Squares (OLS) regression. The goal is to examine the extent to which factors such as age, BMI, smoking status, alcohol use, physical activity, income, sex, and antihypertensive medication use influence SBP. The classical linear regression assumptions (BLUE conditions) are also evaluated whether they hold or not.

$$\begin{aligned}\text{SBP} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{BMI} + \beta_3 \cdot \text{Smoking\_Status} + \beta_4 \cdot \text{Alcohol\_Use} \\ + \beta_5 \cdot \text{Physically\_Active} + \beta_6 \cdot \text{Income} + \beta_7 \cdot \text{Sex} + \beta_8 \cdot \text{Hypertensive\_Meds} + \varepsilon\end{aligned}$$

Hypothesis:

- $H_0 : \beta_j = 0$ (No significant effect of the independent variable on SBP)
- $H_A : \beta_j \neq 0$ (Significant effect of the independent variable on SBP)

2 Data and Variables

Data is sourced from Wave 5 of the Add Health dataset Harris [2025], merged across multiple modules (pwave5, pcardio5, pdemo5, and panthro5) using the respondent identifier AID. The final cleaned dataset includes the following variables:

- **sbp**: Systolic blood pressure (outcome variable)
- **age**: Age in years
- **bmi**: Body Mass Index
- **smoking_status**: Current smoking status (binary)
- **alcohol_use**: Alcohol use (binary)
- **physically_active**: Engaged in physical activity for more than 2 days in the past 7 days (binary)
- **income**: Annual income (categorical)
- **sex**: Male = 1, Female = 0
- **hypertensive_meds**: Use of antihypertensive medication (binary)

Correlation Matrix:

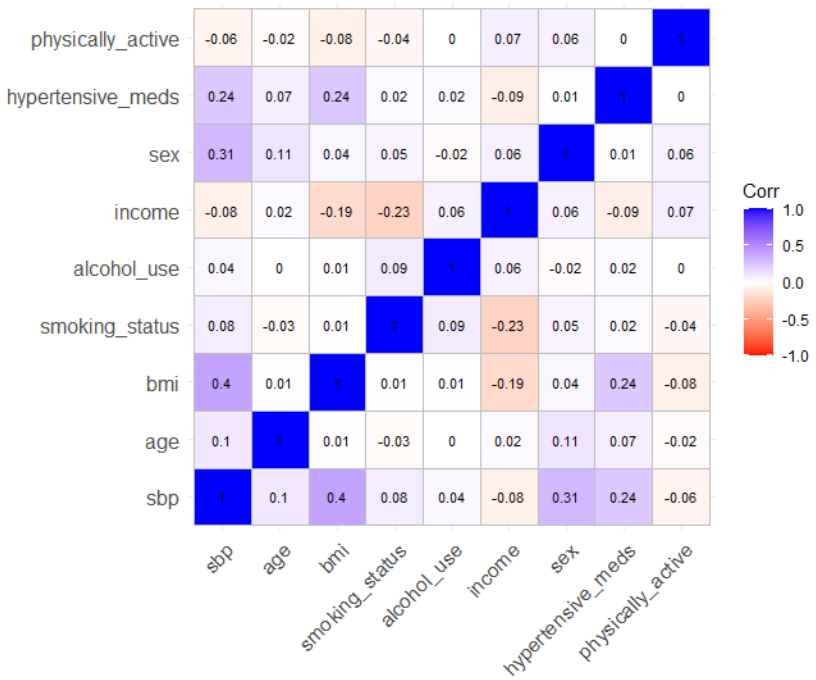

Figure 1: Correlation matrix of variables.

Outliers were removed from continuous variables: sbp, age, bmi, and income, using the 1.5×IQR rule. Variables with invalid/missing codes were also excluded. 1278 observations of 1398 total observations were retained after outlier removal.

Histograms and Boxplots (Before and After Outlier Removal):

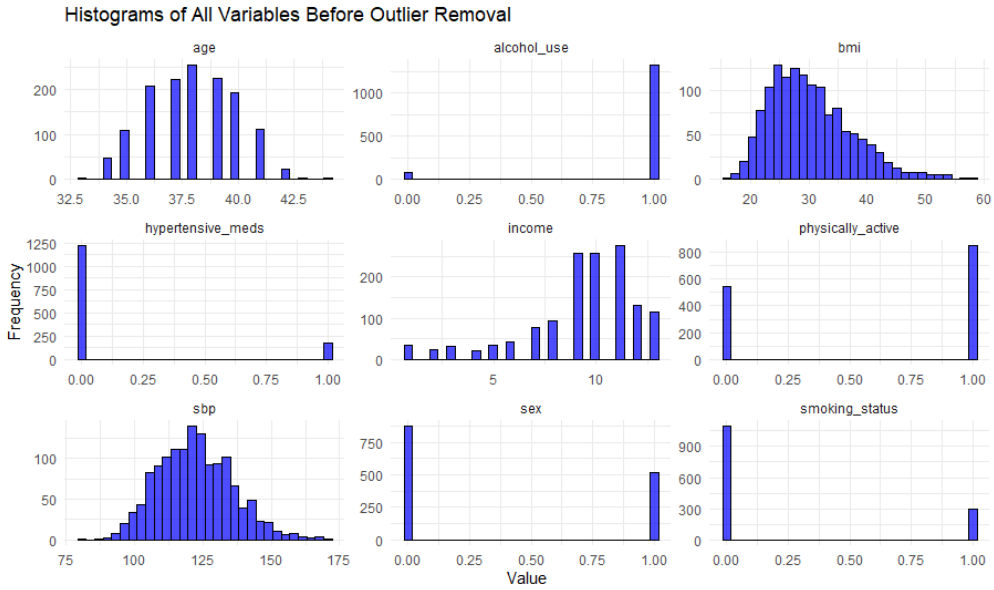

Figure 2: Histograms of variables before outlier removal.

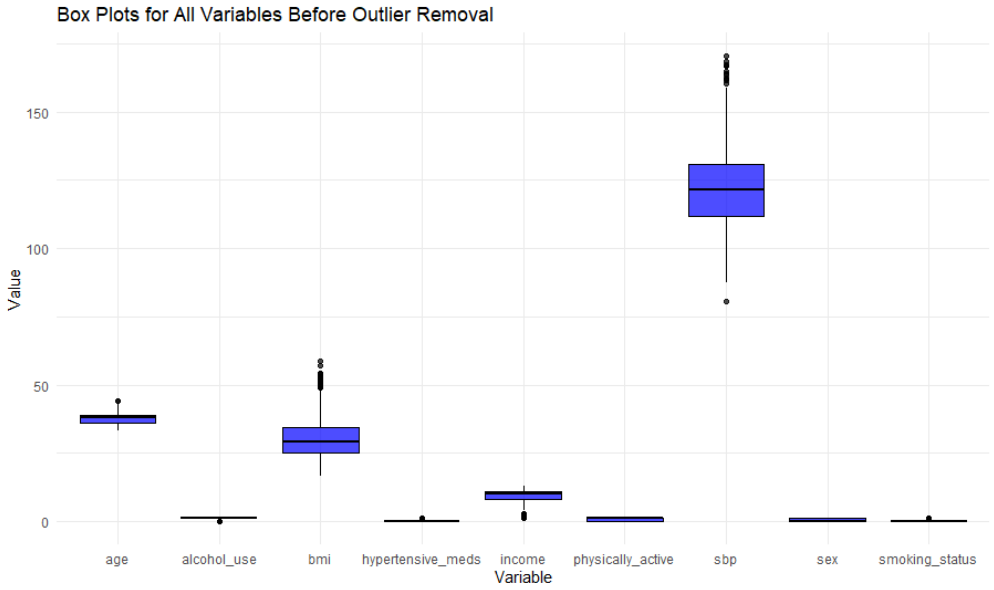

Figure 3: Boxplots of variables before outlier removal.

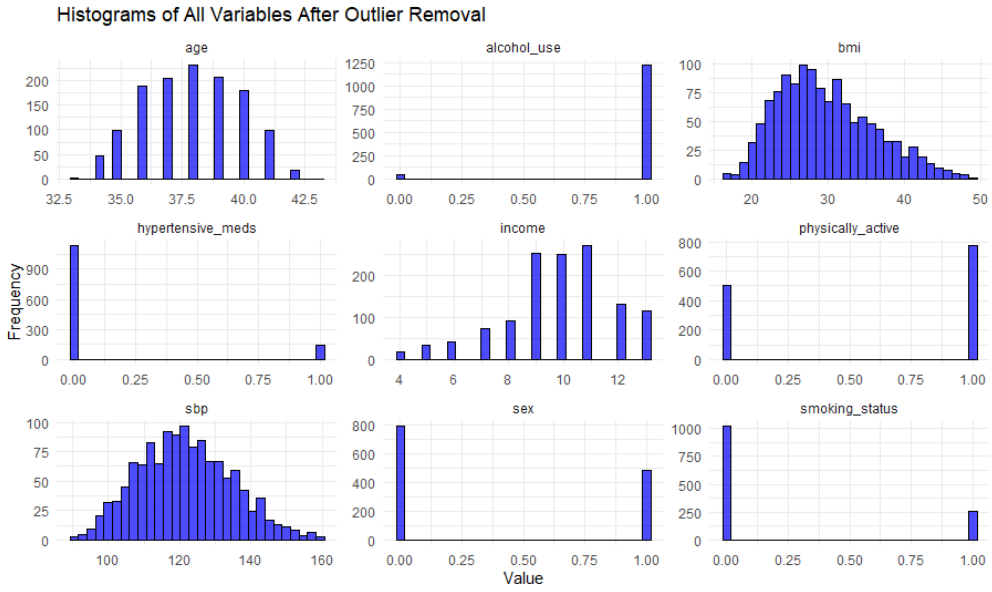

Figure 4: Histograms of variables after outlier removal.

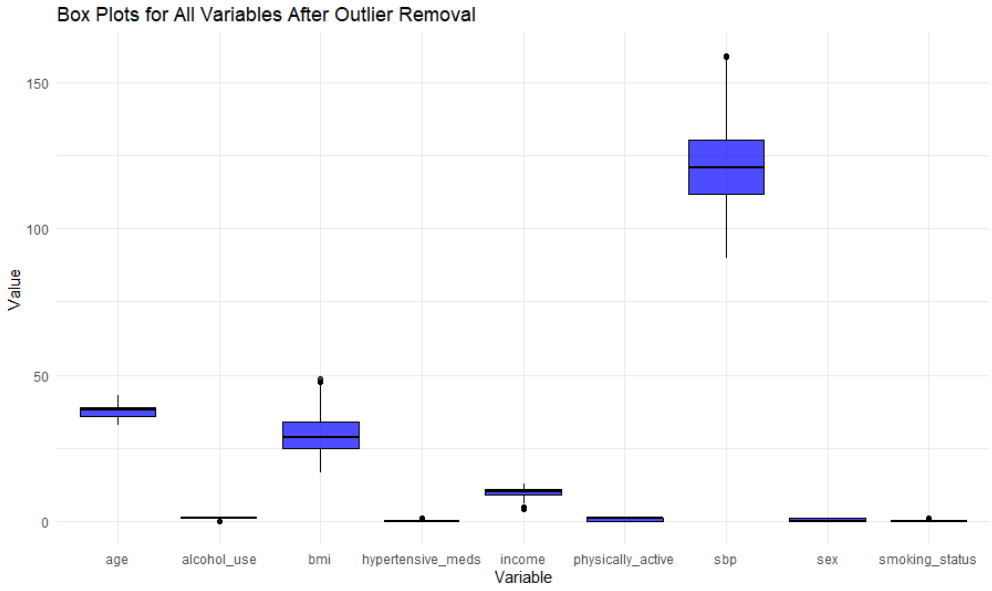

Figure 5: Boxplots of variables after outlier removal.

3 Methods

An OLS regression model was fitted to predict SBP using the predictors. Diagnostic tests were conducted to assess the following assumptions:

- Linearity
- No perfect multicollinearity (VIF)

- Homoscedasticity (Breusch-Pagan, GQ tests)

Robust standard errors were used to account for any potential heteroskedasticity.

4 Results

4.1 OLS Regression Summary

The OLS regression model assessing the predictors of systolic blood pressure (SBP) yielded the following results:

Table 1: OLS Regression Coefficients

Variable	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	82.38875	6.75855	12.190	$< 2e$-16	***
age	0.34137	0.16272	2.098	0.0361	*
bmi	0.70961	0.05193	13.665	$< 2e$-16	***
smoking_status	1.63195	0.81369	2.006	0.0451	*
alcohol_use	2.43121	1.59566	1.524	0.1278	
physically_active	-1.14048	0.64414	-1.771	0.0769	.
income	-0.05558	0.16368	-0.340	0.7342	
sex	7.95409	0.65346	12.172	$< 2e$-16	***
hypertensive_meds	6.12655	1.03005	5.948	$< 3.51e$-09	***

Table 2: OLS Regression Residuals

Min	1Q	Median	3Q	Max
-33.013	-7.585	-0.683	6.912	36.512

- - -
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.17 on 1269 degrees of freedom
Multiple R-squared: 0.2765,
Adjusted R-squared: 0.2719
F-statistic: 60.62 on 8 and 1269 DF,
p-value: $< 2.2e$-16

- Age, BMI, smoking status, sex, and use of hypertensive medication are statistically significant predictors of SBP at 5% confidence.
- Sex has a strong positive association, with males having 7.95 mmHg higher SBP on average.

- Use of hypertensive meds is associated with a 6.12 mmHg higher SBP, possibly reflecting confounding by indication (i.e., those on meds likely had high BP to begin with).
- Physical activity and alcohol use are marginally significant or non-significant.
- Income does not show a significant association with SBP in this model.
- The model explains approximately 27.2% of the variability in SBP, and the overall model is statistically significant.

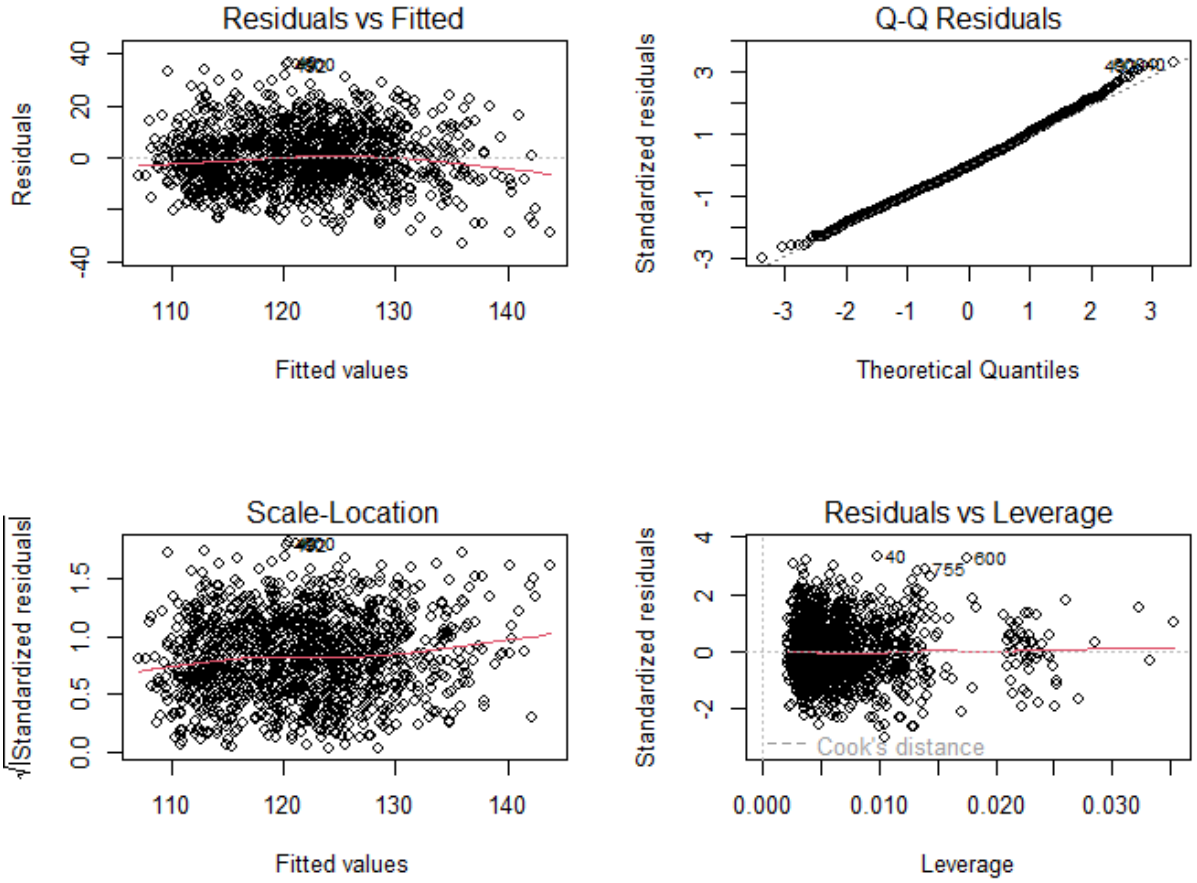

Figure 6: Diagnostic plots for OLS regression.

4.2 RESET Test for Model Specification

RESET test

RESET = 13.362, df1 = 2, df2 = 1267, p = 1.808e-06

The RESET test yields a statistically significant result ($p < 0.001$), indicating evidence of model misspecification. This suggests that the functional form of the model is incorrect or that important variables or interaction terms are missing.

4.3 Multicollinearity (VIF)

Table 3: OLS Regression Coefficients

Variable	VIF
age	1.020
bmi	1.104
smoking_status	1.077
alcohol_use	1.018
physically_active	1.015
income	1.119
sex	1.028
hypertensive_meds	1.067

All Variance Inflation Factor (VIF) values are well below the conventional threshold of 5 (or even 10), suggesting that multicollinearity is not a concern in this model.

4.4 Homoscedasticity Tests

Studentized Breusch-Pagan test

BP = 27.053, df = 8, p-value = 0.0006924

White Test - Studentized Breusch-Pagan test

BP = 13.62, df = 2, p-value = 0.001103

Goldfeld-Quandt test

GQ = 1.0064, df1 = 630, df2 = 630, p-value = 0.4682
alternative hypothesis: variance increases from segment 1 to 2

- The Breusch-Pagan test and White test, both indicate heteroscedasticity, i.e., non-constant variance of residuals, which violates the homoscedasticity assumption of OLS.
- The Goldfeld-Quandt test, however, does not support heteroscedasticity ($p = 0.47$).
- Given the strong evidence from the BP tests, robust standard errors are warranted for inference.

4.5 Robust Standard Errors

Table 4: OLS with Robust Standard Errors (HC1)

Variable	Estimate	Robust SE	t value	Pr(>|t|)	
Intercept	82.388747	6.877477	11.9795	$< 2.2e$-16	***
age	0.341367	0.163326	2.0901	0.03681	*
bmi	0.709614	0.054152	13.1042	$< 2.2e$-16	***
smoking_status	1.631952	0.825368	1.9772	0.04823	*
alcohol_use	2.431213	1.454972	1.6710	0.09497	.
physically_active	-1.140481	0.650667	-1.7528	0.07988	.
income	-0.055583	0.164091	-0.3387	0.73487	
sex	7.954094	0.654282	12.1570	$< 2e$-16	***
hypertensive_meds	6.126548	1.249279	4.9041	$1.06e$-6	***

- - -Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Using heteroscedasticity-consistent standard errors (HC1):

- Age ($p = 0.037$), BMI ($p < 0.001$), Smoking Status ($p = 0.048$), Sex ($p < 0.001$), and Hypertensive Medication Use ($p < 0.001$) remain statistically significant predictors of SBP.
- Alcohol Use and Physical Activity are marginally significant ($p < 0.10$ and < 0.08).
- Income is not statistically significant ($p = 0.73$).

4.6 Exogeneity

While the ordinary least squares (OLS) regression model provides insight into the associations between systolic blood pressure and various predictors (age, BMI, smoking status, etc.), it operates under the assumption that all explanatory variables are exogenous, that is, uncorrelated with the error term. However, in real-world observational data, this assumption is often challenged. Several unmeasured or omitted variables could influence both systolic blood pressure and one or more of the predictors in the model, potentially leading to endogeneity bias. Some examples of such variables include dietary habits, stress and mental health, genetic factors, ethnicity, other illness, and access to healthcare and healthcare education. These omitted or mismeasured variables may be absorbed into the error term, violating the exogeneity assumption and leading to biased and inconsistent coefficient estimates.

Although endogeneity is a critical issue in causal inference, addressing it requires approaches that demand additional assumptions, data, or modeling complexity. The issue of endogeneity is therefore acknowledged as a limitation but is not resolved within the analytical framework used here.

5 Discussion

- BMI shows the strongest association with systolic blood pressure, consistent with the well-documented link between adiposity and hypertension.
- Sex is a significant predictor, with males exhibiting higher SBP than females.
- Hypertensive medication use is associated with higher SBP, possibly due to indication bias, that is, individuals on medication may already have higher blood pressure.
- Age, smoking, and potentially alcohol use and physical activity also influence SBP, though with smaller effect sizes.
- Income, surprisingly, was not a significant predictor in this sample, possibly due to measurement limitations or unaccounted socioeconomic confounders.
- Despite the model explaining about 27% of the variance in SBP (Adjusted $R^2 = 0.272$), model specification issues (RESET test) and heteroscedasticity highlight areas for refinement.
- Recommendations include exploring non-linear terms, and interaction effects between variables.

6 Conclusion

This study identifies body mass index (BMI) as the most influential factor associated with systolic blood pressure (SBP), confirming existing evidence linking adiposity to elevated blood pressure. Sex also plays a significant role, with males exhibiting higher SBP than females. Use of antihypertensive medication correlates with higher SBP, likely due to indication bias, where individuals already diagnosed with hypertension are more likely to be on medication. Additional factors such as age, smoking status, alcohol use, and physical activity show modest associations. Contrary to expectations, income was not a significant predictor of SBP in this sample, potentially reflecting measurement limitations or omitted socioeconomic confounders. While the model explains a moderate proportion of the variance in SBP (Adjusted $R^2 = 0.272$), diagnostic tests suggest issues with model specification and heteroscedasticity. Future research should consider incorporating non-linear modeling and interactions among the variables to better capture the complex determinants of SBP.

References

Beverly H. Brummett, Michael A. Babyak, Ilene C. Siegler, Michael Shanahan, Kathleen Mullan Harris, Glen H. Elder, and Redford B. Williams. Systolic blood pressure, socioeconomic status, and biobehavioral risk factors in a nationally representative us young adult sample. *Hypertension*, 58(2):161–166, 2011. doi: 10.1161/HYPERTENSIONAHA.111.171272. URL https://www.ahajournals.org/doi/abs/10.1161/HYPERTENSIONAHA.111.171272.

Beverly H Brummett, Michael B Babyak, IIene C Siegler, Richard Surwit, Anastasia Georgiades, Stephen H Boyle, and Redford B Williams. Systolic blood pressure and adiposity: Examination by race and gender in a nationally representative sample of young adults. *American Journal of Hypertension*, 25(2):140–144, 02 2012. ISSN 0895-7061. doi: 10.1038/ajh.2011.177. URL https://doi.org/10.1038/ajh.2011.177.

Beverly H. Brummett, Michael A. Babyak, Rong Jiang, Kim M. Huffman, William E. Kraus, Abanish Singh, Elizabeth R. Hauser, Ilene C. Siegler, and Redford B. Williams. Systolic blood pressure and socioeconomic status in a large multi-study population. *SSM - Population Health*, 9:100498, 2019. ISSN 2352-8273. doi: https://doi.org/10.1016/j.ssmph.2019.100498. URL https://www.sciencedirect.com/science/article/pii/S2352827319301892.

Helen M Colhoun, Harry Hemingway, and NR Poulter. Socio-economic status and blood pressure: an overview analysis. *Journal of human hypertension*, 12(2):91–110, 1998.

Carol A. Ford, James M. Nonnemaker, and Kathleen E. Wirth. The influence of adolescent body mass index, physical activity, and tobacco use on blood pressure and cholesterol in young adulthood. *Journal of Adolescent Health*, 43(6):576–583, 2008. ISSN 1054-139X. doi: https://doi.org/10.1016/j.jadohealth.2008.06.010. URL https://www.sciencedirect.com/science/article/pii/S1054139X08002632.

Kathleen Mullan Harris. National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave V, 2016-2018, 2025. URL https://doi.org/10.15139/S3/ZYRZ5J.

Jun Li, Rui-Hua Feng, Yan-Na Mao, Xiao-Wan Wang, and Ning-Ning Wang. Gender-specific differences in associations between economic status and systolic blood pressure or diastolic blood pressure. *Chinese Medical Journal*, 133(14):1722–1724, 2020. doi: 10.1097/CM9.0000000000000953. URL https://mednexus.org/doi/abs/10.1097/CM9.0000000000000953.

Anna Zamora-Kapoor, Luciana E Hebert, Morgan Montañez, Dedra Buchwald, and Ka'imi Sinclair. Risk factors in adolescence for the development of elevated blood pressure and hypertension in american indian and alaskan native adults. *Journal of immigrant and minority health*, 23(4):717–724, 2021.