Analyzing Effect of Weight on Blood Pressure

Lab 2: Datasci 203

Sonia Song, Kenneth Hahn, Mei Qu

Contents

Importance and Context

Hypertension, affecting nearly one-third of adults worldwide, is a leading risk factor for heart disease and stroke. Despite its prevalence, the exact causes of hypertension are not fully understood, and various factors are thought to contribute to its development. One such factor that warrants investigation is body weight. Our analysis seeks to answer the below research question using statistical methods:

What are the differences in blood pressure levels among individuals with varying body weights?

The answers to this question could provide crucial insights, potentially leading to more effective public health strategies, personalized interventions to manage and prevent hypertension, and overall enhancement of health outcomes and reduction of health disparities related to hypertension.

We will also analyze the effect of smoking/secondary smoke as an additional factor. With the information of how many members of the household smoke, we would like to determine if any exposure to smoking can potentially influence the distribution of blood pressure.

Data and Methodology

We sourced data from the National Health and Nutrition Examination Survey, a survey that combines both interview data along with physical examinations to characterize the prevalence of major diseases. From the survey, we will utilize 2 datasets: the physical examination and the questionnaire datasets from 2013-2014.[1] These will be inner joined on `SEQN` - a unique identifier for each respondent. We will take the blood pressure and body weight from the physical examination and household smoking data from the questionnaire to conduct our regression analysis.

To determine the blood pressure of the respondents, we will use the variables `BPXDI1-BPXDI4` and `BPXSY1-BPXSY4` (`BPXDI` = diastolic blood pressure; `BPXSY` = systolic blood pressure, and the appended number is the number of tests conducted on the respondent, in units of mmHg). First, we will remove any zero or NA values as a reading of zero blood pressure is not possible. Then, because there must only be one target variable to represent the blood pressure in our model, we will take the average of the respective pressures for each of the four tests to smooth out any outliers. Finally, we will calculate a "Mean Arterial Pressure" (or `MAP`, a clinical measure to gauge hypertensive and hypotensive states), as follows (`DP` = Diastolic Pressure; `SP` = Systolic Pressure)[2]: $MAP = DP + \frac{1}{3}(SP - DP)$

As for our features, we will represent the body weight with the BMXWT variable, and remove any weights that are NA. To estimate the relationship between smokers and blood pressure for our secondary model, we will utilize the questionnaire dataset that asks the question "How many people [in your household] smoke cigarettes [...] or any other tobacco product?" in the variable `SMD460`. We categorized any values > 0 as a smoking household and any $= 0$ as a non-smoking household. We also removed values > 777 as those responses are not valid. We chose this question as opposed to other smoking related questions because this not only informs of the impact of an individual's smoking habits on blood pressure, but it may also imply the influence of secondhand smoking.

After filtering out NA values and unrealistic outliers, we ended up with a joined data set of 7350 rows, originally beginning from 9813 rows. Due to the size of this dataset we randomly sampled the data into training and test data sets, with 2215 and 5135 rows respectively. With the test dataset, we will perform OLS regression and evaluate the best fitting model from the following:

(1. Simple Model) $MAP = \beta_0 + \beta_1 weight$

(2. Indicator Variable) $MAP = \beta_0 + \beta_1 weight + \beta_2 smoking\_household$

[1] Centers for Disease Control and Prevention. (2014b). 2013-2014 questionnaire data - continuous NHANES. National Health and Nutrition Examination Survey. https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Questionnaire&Cycle=2013-2014

[2] DeMers, D., & Wachs, D. (2023). Physiology, Mean Arterial Pressure. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK538226/#:~:text=A%20common%20method%20used%20to,%2B%201%2F3(PP)

In order to conduct the OLS regression we satisfy the two large-scale assumptions in that the data must be 1) I.I.D. and 2) a unique BLP must exist. Evaluating the first assumption, we can state that the data is independent as each row represents a different respondent and their `MAP` and `weight` will not influence another respondent's `MAP` or body weight. There is a possibility that the smoking household may not be independent if some of the participants were living in the same household; however, we will assume that it is independent through the sampling method and large quantity of data. We can assume that the data is also identically distributed as it is pulling from the underlying distribution of the U.S. population for all points.

To ensure a BLP exists, the covariance of the features and the outcome must be finite. By reviewing the histograms of `MAP` and body weight in Figure 1 we do not observe any heavy tails, which suggests that the covariance is in fact finite. Also, to categorize whether the BLP is unique we must ensure that there is no collinearity between variables. For our first model, we are only estimating the impact of `weight` on `MAP`, so no collinearity exists. For our second model, reviewing the scatterplot of `weight` and MAP in Figure 2, we can see that there does not appear to be collinearity between `weight` and household smoking, which aligns with our understanding that one's weight should not influence whether the household smokes.

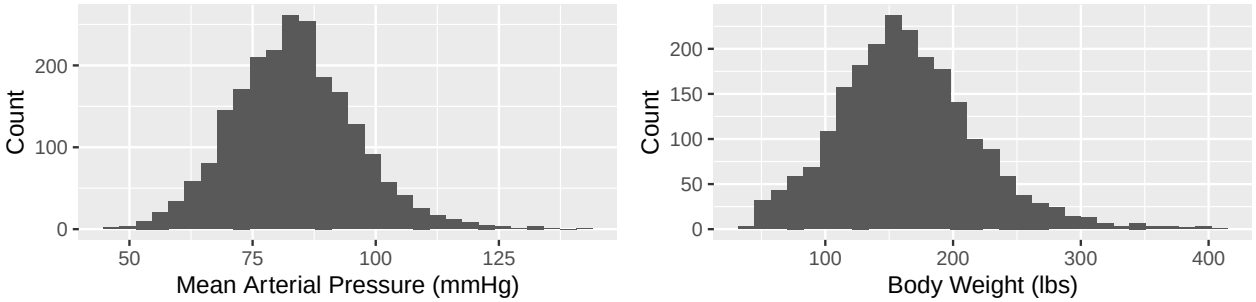

Figure 1: Distributions for MAP (left) and Body Weight (right)

With the large scale assumptions satisfied, we will conduct OLS regression. We will evaluate the models by reviewing the t-test results for the coefficients and we will conduct an F-test to determine if the addition of the smoking household variable adds any significant contribution to MAP. The F-test will be performed in order to evaluate the following null hypothesis and alternative hypothesis:

H_0: *The variable* `smoking_household` *does not contribute to the MAP model*

H_a: *The variable* `smoking_household` *does contribute to the MAP model*

Results

The results of the OLS regression are shown in Table 1. We calculated the robust standard error with the `vcovHC()` method using type `HC3` for both models, assuming that the dataset was heteroscedastic based on the Figure 2 below. The linear regressions for both models suggest that the coefficient for `weight` ($\beta_1 = 0.095$) is statistically significant with p-values effectively equal to zero (<2.2e-16) when conducting a `coeftest()`. The t-test suggests that there is a relationship between `MAP` and `weight` and the coefficient delineates that a 1 lb increase in weight leads to a proportional increase of 0.095 mmHg in MAP for our distribution.

To evaluate the effect of the `smoking_household` indicator variable, we conducted an 'F-test' between our models, where we set the Type I error to be $\alpha = 0.05$. The results of the test indicates that there is no statistically significant difference between the two models (F = 1.39, p = 0.239) in terms of their ability to describe the blood pressure distribution. Specifically, the addition of the indicator variable to the model does not provide an improvement in fit over the simpler model.

We have considered other transformations such as linear-log or log-log models but ultimately decided to use the indicator-variable model on the test dataset to capture additional factors that can influence blood pressure that would be interesting to analyze, see Appendix for various model assumptions. Ultimately, after evaluating the two models, we have decided to proceed with the simple model as our final model. This

Table 1: Comparison of Regression for Mean Arterial Pressure and Weight

	Dependent variable:	
	Mean Arterial Pressure (M.A.P.)	
	Model 1	Model 2 (Smoking Household)
Weight (lbs)	0.095*** (0.003)	0.095*** (0.003)
Smoking Household		0.378 (0.381)
Intercept	68.108*** (0.534)	68.016*** (0.540)
Observations	5,135	5,135
R^2	0.167	0.168
Adjusted R^2	0.167	0.167
Residual Std. Error	11.550 (df = 5133)	11.550 (df = 5132)
F Statistic	1,031.582*** (df = 1; 5133)	516.310*** (df = 2; 5132)
Note:		*p<0.05; **p<0.01; ***p<0.001

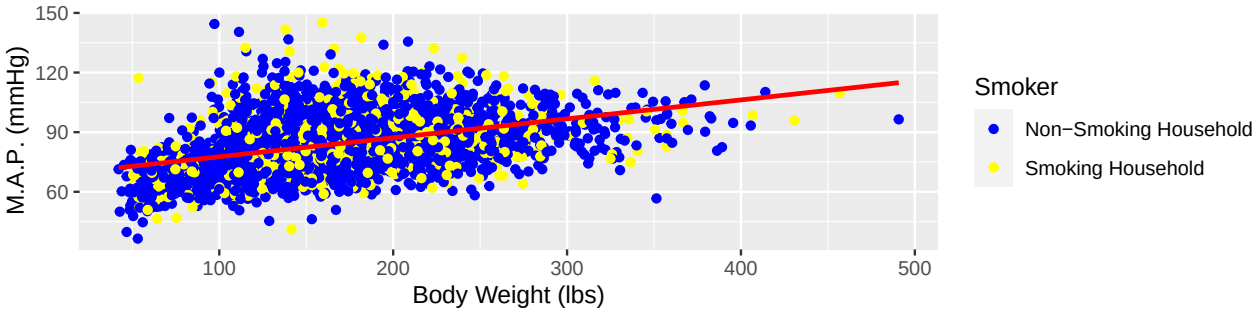

Figure 2: Weight vs. Blood Pressure

decision is based on the lack of significant improvement from incorporating the household smoking variable and on the interpretability of weights implication on `MAP`.

Discussion

For our indicator variable model, we only considered data on the number of people in the household who smoke, resulting in a statistically insignificant improvement over the simple model. Despite multiple studies concluding that smoking is indeed a high-risk factor for hypertension, the broadness of our variable may have limited its effectiveness. Instead of assessing household smoking, a targeted approach that examines the frequency and intensity of just the respondent's smoking behavior could provide more precise insights. Additionally, other transformations are worth considering, one of which is taking the log of body weight. The absolute weight change could have different degrees of influence on a person's health depending on the scale of the original weight. Therefore, taking the log (comparing the percentage of weight change rather than absolute) would stabilize the variable and make the residuals more homoscedastic.

Finally, the questionnaire was conducted in 2013-2014 (10 years ago). We chose this dataset and specific variables given completeness of the data relative to surveys conducted in more recent years. Although updates have been made since then, the time elapsed could potentially lead to data staleness which may not describe the current U.S. population. In conclusion, this study demonstrates a positive relationship between weight and blood pressure. These findings underscore the importance of maintaining a healthy weight as a means of managing blood pressure, hypertension, and promoting overall cardiovascular health.

Appendix

I. Link to our data sources (in CSV format):

https://github.com/hahnkenneth/lab_2_hahn_qu_song/tree/main/data/raw/archive%20(5)

II. Model Specifications we tried to arrive at the final model:

When choosing datasets, we learned that values collected from real-world observations rather than synthetic dataset require more data manipulations but could be more valuable in understanding relationships between variables that inform human behaviors.

Log-Log Model ($log(MAP) = \beta_0 + \beta_1 log(weight) + \epsilon$): Looking at the Residuals-vs-Fitted-values Plot from transforming our simple linear model to a Log-Log model, we saw that the residuals for the Log-Log model are more evenly distributed around zero, indicating less heteroscedasticity and suggesting a better model fit. However, there are clear limitations as well. The model assumes a constant elasticity, which might not hold true across all weight ranges. Additionally, taking the log of blood pressure may not make practical sense as the marginal increase in blood pressure is relatively consistent across the distribution. We could also face overfitting the data if linear regression is already a good fit for the data.

Learning: Explainability needs to be considered in a real-world context to determine suitability of a particular transformation.

Linear-Log Model ($MAP = \beta_0 + \beta_1 log(weight) + \epsilon$):

Learnings: When choosing the independent variable, we learned that it is important to form hypotheses drawn from previous knowledge (e.g. weight may have a linear relationship with blood pressure) when selecting from a large number of factors.

Before arriving at the models used in testing, we tried the following specifications:

1. $MAP = \beta_0 + \beta_1 log(weight) + \epsilon$
2. $log(MAP) = \beta_0 + \beta_1 log(weight) + \epsilon$
3. $MAP = \beta_0 + \beta_1 log(weight) + \beta_2 smokinghousehold + \epsilon$

We did not end up testing 1) as it is too similar from the simple model and we believe that indicator variables should be prioritized given the high probability of unobserved variables influencing predictability of the model. However, if we are not limited by the number of models (2) for testing, it would be worth conducting further research to understand the applicability of this transformation. We ruled out 2) because taking the log of blood pressure does not provide meaningful transformation as mentioned above. Finally, we ruled out 3) because the complexity (combining log and indicator variable) will reduce the interpretability of the final output.

Learnings: Adding variables doesn't necessarily increase the predictive power of the model, even if one may intuitively think that the variable could be correlated with the dependent variable (e.g. household smoke with blood pressure).

III. Residuals-vs-fitted values plot for the simple model (left) and indicator variable model (right):