Math 161B: Applied Statistics and Probability II
Section 01

ANOVA and Regression Analysis of National Health and Nutrition Examination Survey

Group G:
Monica O.
Clarizza A.
Prateetya B.
XinRu L.

May 24, 2021

INTRODUCTION

Blood pressure and weight are important as these are indicators of our health. In addition, there are influences on blood pressure and weight that are based on our lifestyle such as diet and rest. There are also socio-economic influences that are proven to impact blood pressure and weight such as race, gender, income (Adler & Newman, 2002). Data on socio-economic and lifestyle influences on health and health outcomes are collected by the National Health and Nutrition Examination Survey (NHANES), which is a United States program that surveys and assesses the health and nutrition of children and adults.

The prevailing disparities in hypertension and hypertension-related illness are seen in African American communities and the causes of these are unclear (Lackland, 2014). Although the relationship between hypertension and stress is not clear, reputable sources like the Harvard Health Publishing state that reducing stress will help keep the blood pressure down (Harvard Health, 2021). The outcomes from these studies inspired our objectives in this report. We strive to understand the relationship between blood pressure and the predictors, and understand the relationship between a child's health and the socio-economic status of their family. We will investigate these relationships through ANOVA and linear regression models. The observation unit in the models is individual survey respondents, specifically individuals from sampled households in the United States. The population is U.S. citizens, specifically adults and children during the years 2013-2014.

DATA DESCRIPTION

PREDICTORS

Nominal Variables:

Income: Total household income (reported as a range value in dollars)
Gender: Gender of the participant
Race: Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
HH_Food_Worries: Household Worried run out of food
HH_Food_Security_Year: Household food security category for last 12 months
Child_Food_Security_Year: Child food security category for last 12 months

Continuous Variables:

Income_to_Poverty: A ratio of family income to poverty guidelines
Age(y): Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age
Age(m): Age in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam or screening if not examined

RESPONSE
BP_systolic: Systolic Blood Pressure Levels (3rd rdg) mm Hg
Weight: Weight kg

Model 1: Two Way ANOVA with Interaction

Variables of Interest:

- Response: Systolic Blood Pressure
- Factors:
  - Income (*under $20,000 , $20,000-$75,000 , $75,000-$99,999, $100,000+*)
  - Race (*Mexican-American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other Race (Including Multi-Racial*))
  - Income * Race

Hypotheses:

- Hypothesis 1:
  H_0: $\gamma ij = 0$ for all i,j (where i = income, j = race)
  The effect of income on the average systolic blood pressure levels of Americans does not depend on race during the years 2013-2014.
  H_a: At least one $\gamma ij \neq 0$
  The effect of income on the average systolic blood pressure levels of Americans depends on race during the years 2013-2014.

- Hypothesis 2:
  H_0: $\alpha_1 = ... = \alpha_i = 0$
  The average systolic blood pressure levels of Americans does not depend on income during the years 2013-2014.

  H_a: At least one $\alpha_i \neq 0$
  The average systolic blood pressure levels of Americans depend on income during the years 2013-2014.

- Hypothesis 3:
  H_0: $\beta_1 = ... = \beta_6 = 0$
  Race does not affect the average systolic blood pressure levels of Americans during the years 2013-2014.

  H_a: At least one $\beta_j \neq 0$
  Race affects the average systolic blood pressure levels of Americans during the years 2013-2014.

Model:

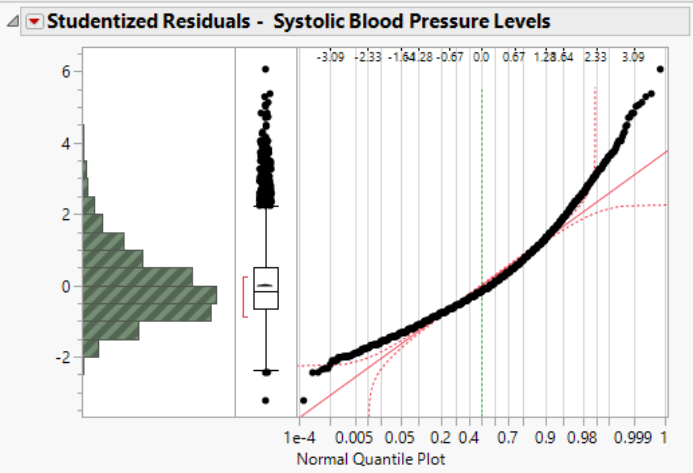
The general (interaction) ANOVA model is:

$$X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{jk} \quad \text{where } i = 1,..,4,\ j = 1,..,6,\ k = 1,..,K$$

The assumption of this model is that ϵ_{ijk} are independent and normally distributed with mean zero and common variance σ^2

Model Checking:

Figure 1a_1: QQ Plot for the Studentized Residuals



The QQ plot in the above figure (right side) displays the normality quantile plots for the studentized residuals of the systolic blood pressure levels in this model. The histogram of the studentized residuals of the systolic blood pressure levels (left side) shows that its distribution is approximately bell-shaped and is centered at mean 0. Also, the normality quantile line does not stray from a straight line and there are over 6,000 observations in this model. Hence, the normality assumption is satisfied.

Figure 1a_2: Studentized Residuals Plot of systolic blood pressure against Race and Income_label

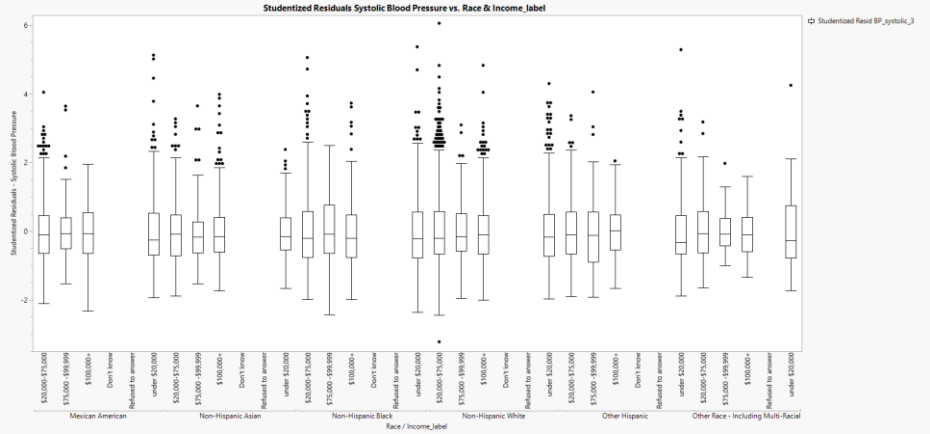
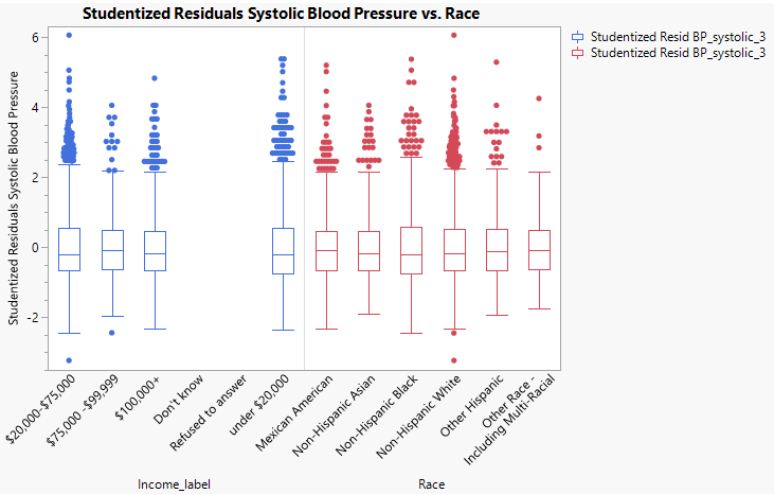


Figure 1a_3: Studentized Residuals of systolic blood pressure for the Factors Race and Income



In the studentized residuals of the treatments (figure 1a_2) and the individual factors from the model (figure 1a_3), we can see from the box plot of the residuals across all groups in the model that the mean is about 0 and the variance is common. Hence, the assumptions of this model have been satisfied.

Results:

Table 1a: Summary of Fit for Model 1

Summary of Fit

Root Mean Square Error	17.89108
Mean of Response	117.8918
Observations (or Sum Wgts)	6838

The observations were reduced to 6838 in this model since survey participants who did not have an income listed were excluded from the calculations for this model. Specifically, participants who had the income label 'Don't Know, 'Refused to Answer', or 'Missing' were excluded.

Table 1b: Analysis of Variance.

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	23	45990.1	1999.57	6.2469
Error	6814	2181097.8	320.09	**Prob > F**
C. Total	6837	2227087.9		<.0001*

Using a significance level of 0.05, there is sufficient evidence to conclude that at least two groups of the average systolic blood pressure levels vary across all treatments in the model.

Table 1c: Effect Tests.

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Race	5	5	19408.678	12.1270	<.0001*
Income_label	3	3	5471.565	5.6979	0.0007*
Income_label*Race	15	15	6732.502	1.4022	0.1362

The income and race factors in the model were statistically significant in this model since boths of its p-values fell below the significance level of 0.05.

Conclusion:

Using the significance level of 0.05, the null hypothesis for the interaction effect of race and income on the average systolic blood pressure is retained (Hypothesis 1) since its p-value exceeded the significance level, while the null hypotheses for the income's effect on the average systolic blood pressure levels (Hypothesis 3) and the effect of race on the average systolic blood pressure levels (Hypothesis 2) have been rejected since its p-values fall below the significance level. We have sufficient evidence to conclude that the effect of income on the average systolic blood pressure does not depend on race and income, the average systolic blood pressure levels depend on race, and the average systolic blood pressure levels depend on income.

Essentially, income and race significantly affected the average systolic blood pressure levels individually, but its combined effect on the average systolic blood pressure levels are insignificant. Since the null hypotheses for the effect of income on average systolic blood pressure levels and for the effect of race on average systolic blood pressure levels were rejected, it makes sense to conduct Tukey's Post Hoc test to check which income groups and which races have significant differences in their effect on average systolic blood pressure levels of Americans.

Further Analysis (Post-Hoc Test):

Table 1d: Tukey HSD All Pairwise Comparisons for Race.

Quantile = 2.85056 , Adjusted DF = 6814.0 , Adjustment = Tukey-Kramer

All Pairwise Differences

Race	-Race	Difference	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Mexican American	Non-Hispanic Asian	-1.66444	1.085275	-1.53	0.6424	-4.75808	1.4292
Mexican American	Non-Hispanic Black	-5.97854	0.974118	-6.14	<.0001*	-8.75532	-3.2018
Mexican American	Non-Hispanic White	-3.67835	0.862623	-4.26	0.0003*	-6.13731	-1.2194

Mexican American	Other Hispanic	-1.68917	1.147973	-1.47	0.6826	-4.96154	1.5832
Mexican American	Other Race - Including Multi-Racial	1.83897	1.535414	1.20	0.8382	-2.53782	6.2158
Non-Hispanic Asian	Non-Hispanic Black	-4.31410	0.986399	-4.37	0.0002*	-7.12589	-1.5023
Non-Hispanic Asian	Non-Hispanic White	-2.01392	0.876468	-2.30	0.1950	-4.51234	0.4845
Non-Hispanic Asian	Other Hispanic	-0.02474	1.158412	-0.02	1.0000	-3.32686	3.2774
Non-Hispanic Asian	Other Race - Including Multi-Racial	3.50341	1.543235	2.27	0.2065	-0.89568	7.9025
Non-Hispanic Black	Non-Hispanic White	2.30018	0.734357	3.13	0.0215*	0.20685	4.3935
Non-Hispanic Black	Other Hispanic	4.28936	1.054990	4.07	0.0007*	1.28205	7.2967
Non-Hispanic Black	Other Race - Including Multi-Racial	7.81751	1.467194	5.33	<.0001*	3.63518	11.9998
Non-Hispanic White	Other Hispanic	1.98918	0.953004	2.09	0.2940	-0.72742	4.7058
Non-Hispanic White	Other Race - Including Multi-Racial	5.51732	1.395662	3.95	0.0011*	1.53890	9.4957
Other Hispanic	Other Race - Including Multi-Racial	3.52814	1.587953	2.22	0.2277	-0.99841	8.0547

After running the Post-Hoc tests for differences between races in the model, the following pairs had significant differences: Mexican-Americans and Black people, Mexican-Americans and White people, Asian people and Black people, Black and White, Black people and other Hispanic people, Black people and other races, and Whites and other races. In general, the post-hoc test revealed that some minorities had significant differences in the average systolic blood

pressure levels. Also, there are significant differences in the average systolic blood pressure levels between some minorities and whites such as Mexican-American people and White people.

Table 1e: Tukey HSD All Pairwise Comparisons for Income.

All Pairwise Differences

Income_label	-Income_label	Difference	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
$20,000-$75,000	$75,000 - $99,999	1.64999	0.992955	1.66	0.3442	-0.90157	4.20155
$20,000-$75,000	$100,000+	2.53626	0.803657	3.16	0.0087*	0.47113	4.60139
$20,000-$75,000	under $20,000	-1.00631	0.759698	-1.32	0.5472	-2.95848	0.94586
$75,000 - $99,999	$100,000+	0.88627	1.143120	0.78	0.8657	-2.05117	3.82371
$75,000 - $99,999	under $20,000	-2.65629	1.112655	-2.39	0.0796	-5.51545	0.20286
$100,000+	under $20,000	-3.54257	0.947579	-3.74	0.0011*	-5.97753	-1.10760

After running the Post-Hoc tests for differences between income groups in the model, the following pairs had significant differences: $20,000-$75,000 and $100,000+, and under $20,000 and $100,000+. In general, the post-hoc test revealed that there are significant differences in the average systolic blood pressure levels between Americans who have high income and those who have low or middle income.

Model 2: ANOVA

Variables of Interest:

- HH_Food_Worries
- HH_Food_Security_Year
- Blood pressure (Systolic)

Hypothesis:

H_0: The average systolic Blood pressure levels are not affected by food worries and food security in the United States during the years 2013-2014.

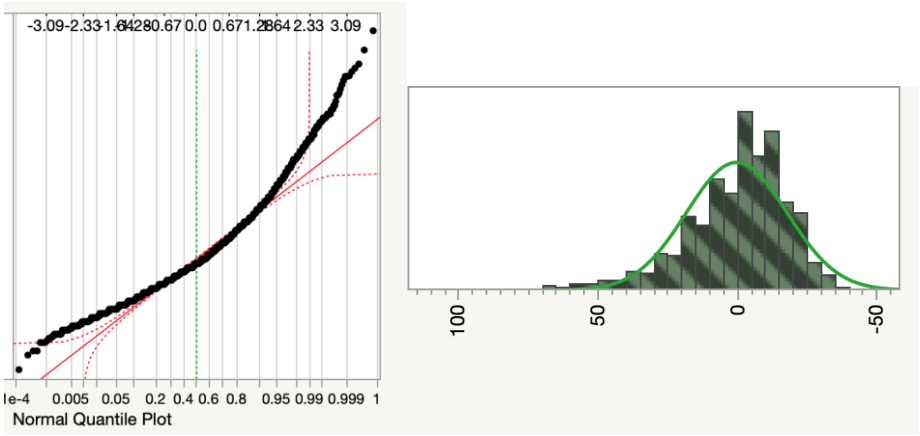
H_a: The average systolic Blood pressure levels are affected by food worries and food security in the United States during the years 2013-2014.

Commented [1]: these hypotheses seem to refer to the interaction between food security and food worries, but I don't see a p-value for the interaction.

Maybe there should be two individual hypotheses for food security and food worries based on the results from the effects table?

Assumptions: The observations are independently, normally distributed

Figure 2a: Normal Quantile Plot (Left) and Normal Distribution Curve (Right). Although the normal quantile plot for ANOVA Model 2 shows a slight deviation from the normal line indicates a normal distribution, the histogram on the right shows an approximately normal distribution centered at 0. We have enough data points in our model to apply the Central Limit Theorem.



Model - The general (interaction) two-way / additive ANOVA model is:

$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{jk}$ where i = 1,..,I, j = 1,..,J, k = 1,..,K

The assumptions of this model are:

- ϵ_{ijk} are independent and normally distributed with mean zero and common variance σ^2

Table 2a: Analysis of Variance for ANOVA Model 2. Looking at the p-value suggests that with a significance level of 0.05, we reject the null hypothesis that blood pressure is affected by food worries and food security in the United States during the years 2013-2014.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	Prob>|t| p-value
Model	5	8806.2	1761.24	5.3952	0.0001
Error	7319	2389251.7	326.45		
C. Total	7324	2398057.9			

Table 2b: Effect Tests for ANOVA Model 2. The only significant factor in determining the average systolic blood pressure of the adult respondents is household food worries (HH_Food_Worries).

Source	Nparm	Degrees of Freedom	Sum of Squares	F Ratio	Prob>|t| p-value
HH_Food_Worries	2	2	2602.8976	3.9867	0.0186*
HH_Food_Security_Year	3	3	298.7743	0.3051	0.8217

Results:
According to the Effect Tests table above, using a significance level of 0.05, we have sufficient evidence to reject the ANOVA model on Food Worries. We also have enough evidence to accept the ANOVA model on Food Security. Therefore, we can conclude that the blood pressure is not affected by Household Food Worries whereas the blood pressure is affected by household food security.

ANOVA Model 3: Multi-factor ANOVA

Variables of Interest:

- Income (*under $20,200* , *$20,000-$75,000*, *$75,000-$99,999*, *$100,000+*) I =4
- Child food security (*Low*, *Marginal*, *Full/Marginal*) J =3
- Insurance (*Yes*, *No*) K = 2
- Weight

Hypothesis:
H_0: Average weight of children ages 0-24 months old is not affected by food worries, household income and insurance enrollment in the United States during the years 2013-2014
H_a: Average weight of children ages 0-24 months old is affected by food worries, household income and insurance enrollment in the United States during the years 2013-2014.

Assumptions: The observations are independently, normally distributed.
Model - The general multi-factor ANOVA model is:

$X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl}$ where i = 1,..,4, j = 1,..,3, k = 1,..,2, l = 1=..L

X_{ijkl} = weight of the l^{th} child income group i, food security j, and insurance category k

The assumptions of this model are:

- ϵ_{ijkl} are independent and normally distributed with mean zero and common variance σ^2

Figure 3a : Student residuals of the data points used in ANOVA Model 3. The studentized residual plot suggests a relatively normal distribution of the residuals and the model assumptions are held despite a couple of outliers.

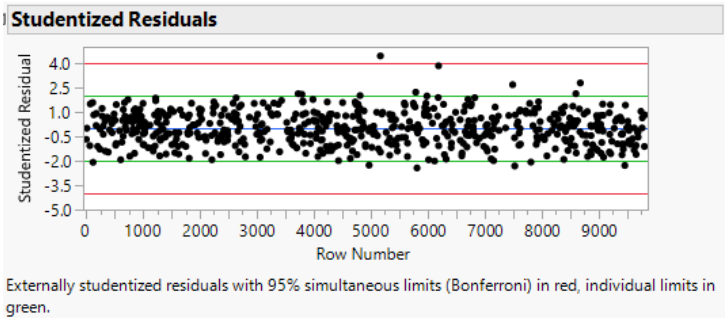
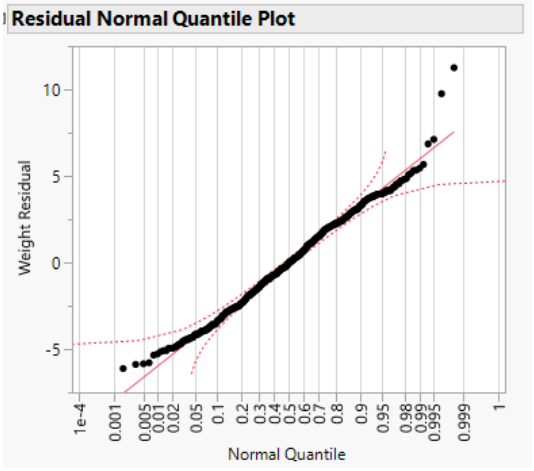


Figure 3b : Normal Quantile plot for the data points used in ANOVA Model 3. The normal quantile plot for model 3 also supports the assumption of residual normality. Although there are a handful of outliers, this is sufficient for our dataset since we have enough data points (600 data points).



Dataset: The dataset used for this multi-factored ANOVA model excludes responses that were labeled as Don't know, Missing, and Refused to answer. This helps focus on those who have given usable information and remove the unwanted effects of responses that provide no meaningful data in the context of this multi-factor ANOVA model.

Results: Using the significance level of 0.05, there is sufficient evidence that Child_Food_Security is significantly related to a toddler's weight and both enrollment in insurance and family income has no significant effect on a toddler's weight.

Table 3a: ANOVA table for model 3. This shows that we have sufficient data to fail to reject the null hypothesis.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	Prob>|t| p-value
Model	5	40.7220	8.14440	1.2273	0.2947
Error	594	3941.7638	6.63597		
C. Total	599	3982.4859			

Table 3b: Effect Tests for ANOVA Model 3. This shows that Child_Food Security_Year is the only significant factor affecting the average weight of a child.

Source	Nparm	Degrees of Freedom	Sum of Squares	F ratio	Prob > F
Child_Food_Security_Year	1	1	34.709072	5.2304	0.0225
Insurance	1	1	0.211507	0.0319	0.8584
Income_label	3	3	6.380056	0.3205	0.8106

There seems to be only one factor (Child_Food_Secrity_Year) that affects the weight of children ages from 0-24 months old. To be sure that there were no dependent effects between the factors of ANOVA Model 3, we ran a one way ANOVA on all the factors. The results for these 3 one way ANOVA models indicate that indeed child food security is in fact (Child_Food_Secrity_Year) is the only factor that affects a child's weight.

Variables of Interest:

- Income (*under $20,200* , *$20,000-$75,000* , *$75,000-$99,999*, *$100,000+*) I =4
- Weight

Hypothesis:

H_0: Average weight of children ages 0-24 months old is not affected by household income in the United States during the years 2013-2014

H_a: Average weight of children ages 0-24 months old is affected by household income in the United States during the years 2013-2014.

Assumptions: The observations are independently, normally distributed.

Model: The general one way ANOVA model is:

$X_{ij} = \mu + \alpha_i + \epsilon_i$ where i = 1,..,4, j = 1,..,J

X_{ij} = weight of the j^{th} child with income group level i

The assumptions of this model are:

- ϵ_{ij} are independent and normally distributed with mean zero and common variance σ^2

Figure 3c: Distribution of Income_label. Although there is not a normal distribution, we have a substantial number of data points (600 data points) to make a one way ANOVA reasonable.

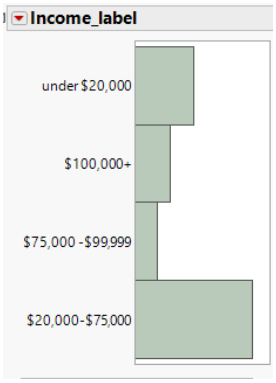


Table 3c: ANOVA table for Income_label factor in ANOVA Model 3. This shows that we have sufficient data to fail to reject the null hypothesis with 95% confidence.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	Prob>|t| p-value
Model	6	5.7857	1.92858	0.2890	0.8333
Error	596	3976.7001	6.67232		
C. Total	599	3982.4558			

Variables of Interest:

- Insurance (*Yes*, *No*) I = 2
- Weight

Hypothesis:

H_0: Average weight of children ages 0-24 months old is not affected by insurance enrollment in the United States during the years 2013-2014

H_a: Average weight of children ages 0-24 months old is affected by insurance enrollment in the United States during the years 2013-2014.

Assumptions: The observations are independently, normally distributed.

Model - The general one way ANOVA model is:

$X_{ij} = \mu + \alpha_i + \epsilon_i$ where i = 1,..,2, j = 1,..,J

X_{ij}= weight of the j^{th} child with Insurance level i

The assumptions of this model are:

- ϵ_{ij} are independent and normally distributed with mean zero and common variance σ

Figure 3d: Distribution of Insurance. Although there is not a normal distribution, we have a substantial number of data points (600 data points) to make a one way ANOVA reasonable.

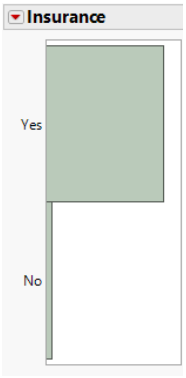


Table 3d: ANOVA table for Insurance factor in ANOVA Model 3.This shows that we have sufficient data to fail to reject the null hypothesis with 95% confidence.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	Prob>|t| p-value
Model	1	0.1447	0.14468	0.0217	0.8829
Error	598	3982.3412	6.65943		
C. Total	599	3982.4858			

Variables of Interest:

- Child food security *(Low, Marginal, Full/Marginal*) I =3
- Weight

Hypothesis:

H_0: Average weight of children ages 0-24 months old is not affected by household income in the United States during the years 2013-2014

H_a: Average weight of children ages 0-24 months old is affected by household income in the United States during the years 2013-2014.

Assumptions: The observations are independently, normally distributed.
Model - The general one way ANOVA model is:

$X_{ij} = \mu + \alpha_i + \epsilon_i$ where i = 1,..,3, j = 1,..,J

X_{ij} = weight of the j^{th} child with Child_Food_Security_Year level i

The assumptions of this model are:

- ϵ_{ij} are independent and normally distributed with mean zero and common variance σ^2

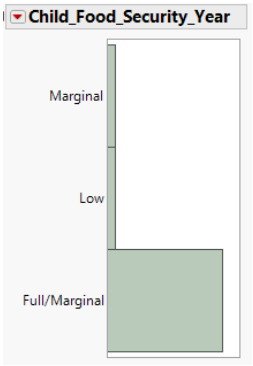


Figure 3e: Distribution of Child_Food_Security_Year. Although there is not a normal distribution, we have a substantial number of data points (600 data points) to make a one way ANOVA reasonable.

Table 3e: ANOVA table for Child_Food_Security_Year factor in ANOVA Model 3. This shows that we have sufficient data to fail to reject the null hypothesis with 95% confidence.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	Prob>|t| p-value
Model	2	45.2490	22.6245	3.4305	0.0330
Error	597	3937.2368	6.5950		
C. Total	599	3982.4858			

Result:

The data from the three individual one way ANOVA models of the factors in ANOVA Model 3 further cements the conclusion that insurance enrollment and income does not affect the average weight of a child between the age 0-24 months old.

Model 4: Multiple Linear Regression

Variables of Interest:

- Blood Pressure (systolic)
- Income**:** Total household income (reported as a range value in dollars)
- Age in years
- Race: Mexican American, Other Hispanic, NonHispanic White, NonHispanic Black, NonHispanic Asian, Other Race
- Gender

The multiple linear regression model we chose is the following:
$\text{Log(Blood Pressure)} = \beta_0 + \beta_1(\textit{Income}) + \beta_2(\textit{Age}) + \beta_3(\textit{Mexican American}) + \beta_4(\textit{Other Hispanic}) + \beta_5(\textit{NH White}) + \beta_6(\textit{NH Black}) + \beta_7(\textit{NH Asian}) + \beta_8(\textit{Gender}) + \varepsilon_i$

Hypothesis:
H_0: Income, Age, Race, and Gender do not have an effect on the average systolic Blood Pressure levels.
$\beta_i = 0,\ i = 1, 2, \ldots, 8$

H_a: Income, Age, Race, and Gender have an effect on the average systolic Blood Pressure levels.
$\beta_i \neq 0,\ i = 1, 2, \ldots, 8$

Checking Assumptions:
We first conduct a global F-test to verify that the average systolic blood pressure depends on at least one of the predictors. With p-value<0.001, we conclude that at least one of the predictors has an effect on an individual's systolic blood pressure.

In this model, we assume that the observations are independent and that the residuals are normally distributed with constant variance and mean zero. We verified these assumptions by looking at the residuals of the dataset. The non-transformed residuals (purple) appeared to violate normality with the curve at one end of the tail. To fix this, we transformed the response variable (green). The tails still curve at the end, but we conclude that considering our large observation size (over 9,000), the observations are close enough to being normal.

Figure 4a, 4b, and 4c: Normal Quantile Plots and Distribution of the Residuals for the Multiple Regression Model. The residuals also appear to be randomly scattered in figure 4c, which

indicates that the systolic blood pressure levels have the same variance. However, the residuals are not quite normal, so we decided to apply a logarithmic transformation to the response.

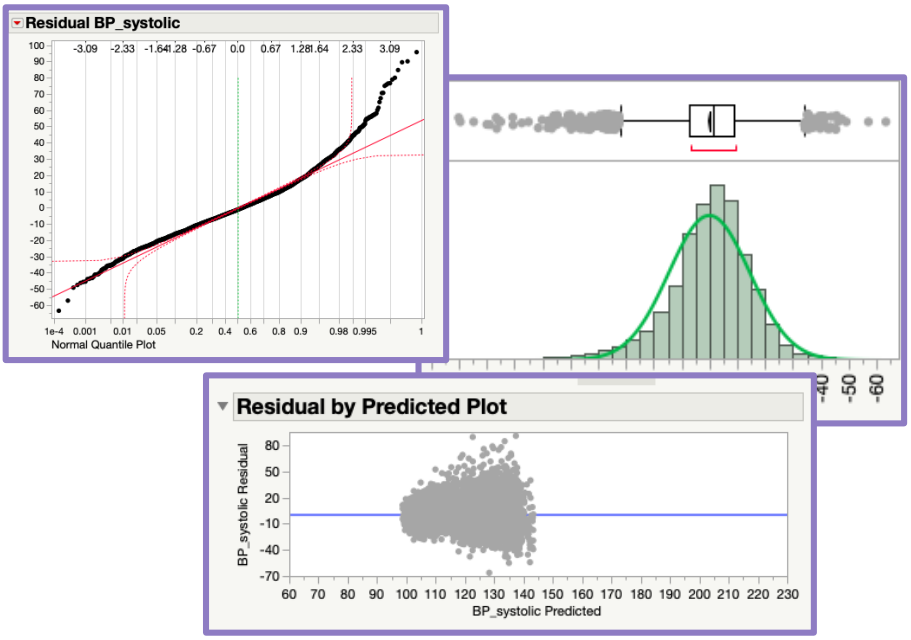


Figure 4d, 4e, and 4f: Normal Quantile Plots and Distribution of the Residuals for the Multiple Regression Model after transformation. The residuals appear to be more normal than Figures 4a, 4b, and 4c. The residuals also appear to be randomly scattered in figure 4f, which indicates that the systolic blood pressure levels have the same variance. Hence, the model assumptions have been satisfied.

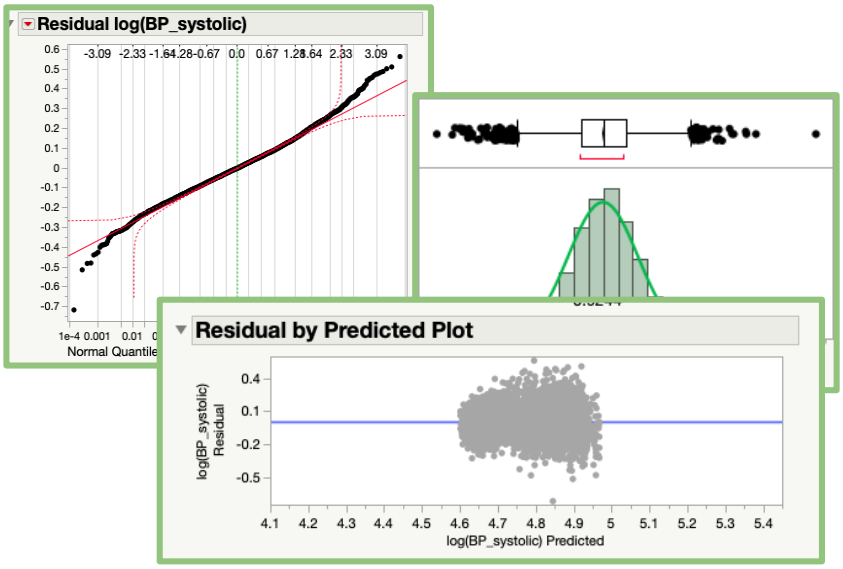


Table 4a: Summary of Fit for Linear Regression Model of the Original Model(Before Variable Selection). This displays the summary of fit for the original model before we made the variable selection model.

R Square	0.338491
R Square Adjusted	0.338009
Root Mean Square Error	14.66923
Mean of Response	117.9397
Observations (or Sum Weights)	6866

Table 4b: Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	100.92911	0.437028	230.94	<.0001*
Income_to_Proverty	-0.525704	0.113248	-4.64	<.0001*
Race[Mexican American]	-0.134497	0.443362	-0.30	0.7616

Race[Non-Hispanic Asian]	-1.112623	0.504011	-2.21	0.0273*
Race[Non-Hispanic Black]	3.0024213	0.384478	7.81	<.0001*
Race[Non-Hispanic White]	-1.690323	0.332966	-5.08	<.0001*
Race[Other Hispanic]	-0.895496	0.536529	-1.67	0.0952
Age(y)	0.4772484	0.008283	57.62	<.0001*
Gender[Female]	-1.680029	0.177311	-9.48	<.0001*

Based on the general additive model, we made a few interesting discoveries. In Figure 4d, we saw that most of the variables have an effect on an individual's blood pressure, with the exception of a few variables in the race category. It was not a huge surprise to see that there is an inverse relationship between income and blood pressure. A person's financial situation affects their stress levels; therefore, income can affect blood pressure as well.

Race has the most interesting results. We found that there is the most statistically significant change in blood pressure when the individual is either black or white. Those that are of a Hispanic descent do not show a statistically significant effect in blood pressure. Whether someone is Asian or not, on the other hand, does have an effect on blood pressure.

Evidence shows that age and gender also affect blood pressure. Older people typically have higher blood pressures. The data also shows that men often have higher blood pressures than women.

Model 5: Variable Selection (Forward/Backward)

We have also chosen to use a Variable Selection Method to further simplify the Multiple Linear Regression Model. With Variable Selection, it intends to select the best subset of predictors in order to explain the model in simplest possible ways. Removing redundant predictors would be one of the ways to simplify the model. Collinearity might also appear if there are many variables in the model since one variable can correlate with the multiple other variables. Prior to using variable selection, there are few criteria that should be met. Such as:

1. Highly adjusted R^2
2. Significant predictors (p-value < 0.05). If p-value > 0.05, remove that predictor.
3. Residual should be normally distributed (i.e, $\mu = 0, \sigma^2 = 1$).

Among all the procedures used for variable selection, we have chosen to use Stepwise Procedure. There are two types of selection in Stepwise Procedure: *Backward Selection* and *Forward Selection.*

Variable of Interest:

- Blood Pressure (systolic)
- Income**:** Income to Poverty Ratio
- Age**:** Age in years
- Race: Mexican American, Other Hispanic, NonHispanic White, NonHispanic Black, NonHispanic Asian, Other Race
- Gender

Checking Assumptions:

Figure 5a: Residual vs Predicted Plot. This displays the Residual vs Predicted plot. Since the residuals appear to be randomly scattered, this verifies the variance assumption. The plot meets the normality assumption because most of the data is around the center of the plot. Also, we have enough data to apply the Central Limit Theorem.

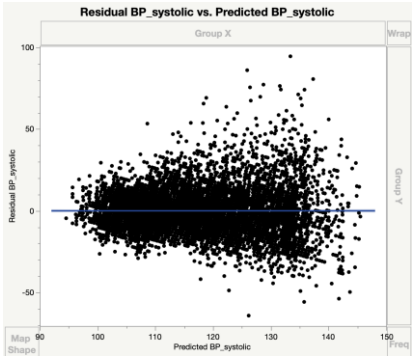


Figure 5b: Normal Quantile Plot (Left) and Normal Distribution (Right) of Backward Selection. This shows the normal quantile plot and the normal distribution curve that supports the assumption of residual normality. The histogram on the right suggests that it is approximately normal distribution with the majority of the data in the center of the curve.

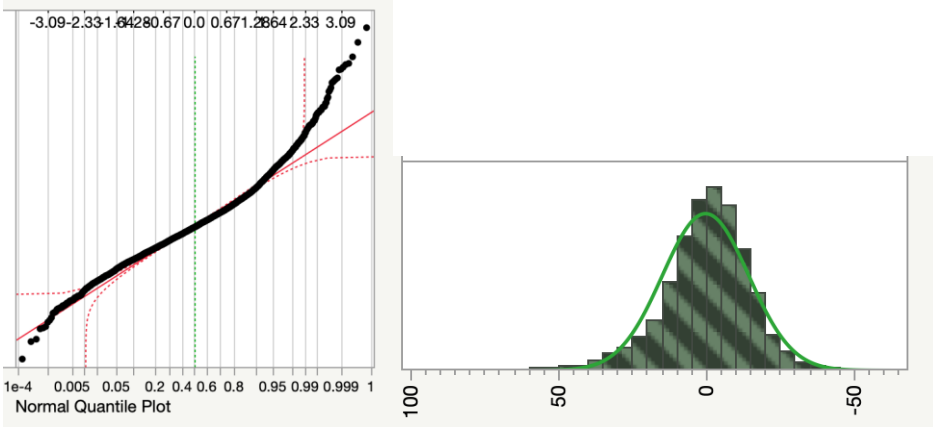


Table 5a: Summary of Fit for linear regression model from Backwards and Forwards Selection. This displays the summary of fit for the linear regression model obtained from the Backward and Forward Selection. The R^2value for the model came out to be 0.345568 with the variable selections. This means that approximately 34.56% of the variation in the average systolic blood pressure levels can be explained by the model. There were just minor changes between the two models (4a and 5a). We have achieved slightly higher R^2value through the variable selection.

R Square	0.345568
R Square Adjusted	0.344805
Root Mean Square Error	14.59374
Mean of Response	117.9397
Observations (or Sum Weights)	6866

Backward Selection

Table 5b: Effect Tests for Backward Selection. Statistically Significant predictor variables shown below in the table with its p-values.

Source	Nparm	Degrees of Freedom	Sum of Squares	F Ratio	Prob > F
Income to Poverty	1	1	5350.64	25.1230	<0.0001
Race	1	1	2425.70	11.3895	0.0007
Age(y)	1	1	702100.30	3296.598	<0.0001
Gender	1	1	11447.96	53.7520	<0.0001
Age(y)*Gender	1	1	9368.12	48.9865	<0.0001
Income to Poverty*Gender	1	1	1599.72	7.5112	0.0061
Race*Age(y)	1	1	6134.30	28.8026	<0.0001

Model: The multiple linear regression that we used for Variable Selection is:

$$\text{Blood Pressure} = \beta_0 + \beta_1(\textit{Income}) + \beta_2(\textit{Age}) + \beta_3(\textit{NH White}) + \beta_4(\textit{NH Black}) + \beta_5(\textit{NH Asian}) + \beta_6(\textit{Mexican American}) + \beta_7(\textit{Other Hispanic}) + \beta_8(\textit{Gender}) + \beta_9(\textit{Income} \times \textit{Age}) + \beta_{10}(\textit{Income} \times \textit{Race}) + \beta_{11}(\textit{Income} \times \textit{Gender}) + \beta_{12}(\textit{Age} \times \textit{Race}) + \beta_{13}(\textit{Age} \times \textit{Gender}) + \beta_{14}(\textit{Gender} \times \textit{Race}) + \varepsilon_i$$

Results:
According to table 5c, all of the variables listed in the table are statistically significant. In other words, we have enough evidence that the variables listed there are the best predictor of the Blood Pressure. We can see that the p-value for all of them is lower than the significance level of 0.05. Compared with the original model, few of the interaction variables have been removed because of nonsignificant factors. Therefore, for the Backward Selection Model, the best possible multiple linear regression model with significant predictors is:

$$\text{Blood Pressure} = \beta_0 + \beta_1(\textit{Income}) + \beta_2(\textit{Age}) + \beta_3(\textit{NH White}) + \beta_4(\textit{NH Black}) + \beta_5(\textit{NH Asian}) + \beta_6(\textit{Mexican American}) + \beta_7(\textit{Other Hispanics}) + \beta_8(\textit{Gender}) + \beta_9(\textit{Age} \times \textit{Gender}) + \beta_{10}(\textit{Income} \times \textit{Gender}) + \beta_{11}(\textit{Race} \times \textit{Age}) + \varepsilon_i$$

Forward Selection:
Table 5c: Effect Tests for Forward Selection shows the variables that have a strong effect on the systolic blood pressure.

Source	Nparm	Degrees of Freedom	Sum of Squares	F Ratio	Prob > F
Income to Poverty	1	1	4838.80	22.7361	<0.0001
Race (Mexican-American, NH Asians & Other Hispanics)	1	1	836.28	3.9294	0.0475
Age(y)	1	1	702389.54	3300.319	<0.0001
Gender	1	1	11616.23	54.5812	<0.0001
Age(y)*Gender	1	1	9504.65	44.6595	<0.0001
Race(NH White & NH Black)*Age(y)	1	1	6279.45	29.5053	<0.0001

Model: The multiple linear regression that we used for Variable Selection is:

$$\text{Blood Pressure} = \beta_0 + \beta_1(\textit{Income}) + \beta_2(\textit{Age}) + \beta_3(\textit{NH White}) + \beta_4(\textit{NH Black}) + \beta_5(\textit{NH Asian}) + \beta_6(\textit{Mexican American}) + \beta_7(\textit{Other Hispanic}) + \beta_8(\textit{Gender}) + \beta_9(\textit{Income} \times \textit{Age}) + \beta_{10}(\textit{Income} \times \textit{Race}) + \beta_{11}(\textit{Income} \times \textit{Gender}) + \beta_{12}(\textit{Age} \times \textit{Race}) + \beta_{13}(\textit{Age} \times \textit{Gender}) + \beta_{14}(\textit{Gender} \times \textit{Race}) + \varepsilon_i$$

Results:

Among all the predictors in the original model shown above, we were able to simplify and choose the significant predictors for the simplified model. The simplified linear regression that we obtained is:

$$\text{Blood Pressure} = \beta_0 + \beta_1(\textit{Income}) + \beta_2(\textit{Age}) + \beta_3(\textit{Mexican American}) + \beta_4(\textit{NH Asian}) + \beta_5(\textit{Other Hispanic}) + \beta_6(\textit{Gender}) + \beta_7(\textit{Age} \times \textit{Gender}) + \beta_8(\textit{NH White} \times \textit{Age}) + \beta_9(\textit{NH Black} \times \textit{Age}) + \varepsilon_i$$

CONCLUSION AND RECOMMENDATION

CONCLUSION

For the Income_label predictors, we had to regroup them as they were overlapping; for example, there is a household income option for *$0 to $ 4,999* and another option for *Under $20,000*. To make it viable for appropriate ANOVA model analysis we had to recode the values of the Income_label predictor variable to *under $20,200* , *$20,000-$75,00*, *$75,000-$99,999*, and *$100,000+*.

Although there is no conclusive research stating that hypertension is an indicator of stress, stress does play a role in a person's blood pressure. Therefore, we are using the systolic blood pressure data as an indicator of health when conducting our analysis. When measuring blood pressure of adults, we have sufficient data that Food Security affects the average blood pressure of adults.

After investigating the Two Way Interaction Model with the factors Race and Income (Model 1), we found that the interaction between Income and Race does not have a significant effect on the average systolic blood pressure levels of Americans. However, Income and Race individually have a significant effect on the average systolic blood pressure. This led us to conduct a Tukey's HSD Test to determine which races and pairs of income groups are significantly different. Based on Tukey's HSD Test for Race, several minorities have significant differences in the average systolic blood pressure levels of Americans between each other. There are also a few minority groups that have significant differences in the average systolic blood pressure levels with white people. Furthermore, Tukey's HSD test for Income reveals that there are significant differences in the average systolic blood pressure levels between those who have a high-income ($100,000+) and those who have low and middle incomes.

We investigated the effect on income, insurance and child food security at home on the average weight of children ages 0 - 24 months old through an ANOVA model (Model 3). We used weight as a proxy for health measurement for children's health as there are no blood pressure measurements for children this young. It is the most reasonable predictor as children struggling with food security are malnourished or underweight (Moradi, 2019). With the children's weight ANOVA model, we are able to conclude that children's weight is dependent on their food security and not their income or insurance enrollment status.

We have also investigated the effects of Income, Age, Gender and Race on the Systolic Blood Pressure through a multiple linear regression model (Model 4). The Race variable in our model consists of multiple races such as: Non Hispanic Whites, Non Hispanic Blacks, Non

Hispanic Asians, Mexican American, and Other Hispanics. The impact of certain races is more significant than other races. We found that Mexican American and the Other Hispanic are less likely to have an effect on Systolic Blood Pressure than other racial groups. The rest of the variables we tested appear to have a statistically significant effect on blood pressure as well.

Additionally, we used the variable selection model (Model 5) to simplify the linear regression model with the best predictor variables. As we used both the Forward and Backward Selection, we obtained the simplified predictors in our model. For the Forward Selection model, we obtained that the best predictor variables are Income, Age, Mexican American, Non Hispanic Asian, Other Hispanic, Gender, interaction between Age and Gender, interaction between NonHispanic Blacks and Age, and interaction between NonHispanic Black and Age. For the Backward Selection model, we resulted with these significant variables: Income, Age, all of the racial groups, Gender, interaction between Age and Gender; interaction between Race and Gender and the interaction between Race and Age.

RECOMMENDATION

After investigating the effect on income, insurance and child food security at home on the average weight of children ages 0 - 24 months old through an ANOVA model (Model 3), we recommend we would need a separate dataset that stated the household income in a quantitative scale and also have the food security rating be quantitative rather than nominal to get a more nuanced model if this effect is to be further investigated. Quantitative variables can allow for a linear regression analysis and use the R squared or R squared adjusted to evaluate the model.

We noticed that the R-squared for the linear regression model (Model 4) is substantially low (33.84%). The predictors of the initial linear regression did not even explain half of the variation in the average systolic blood pressure levels of Americans. Even with the added interaction terms in the backwards and forwards variable selection models (Model 5), the R-squared for both models is 34.56%, which again does not explain even half of the variability on the average systolic blood pressure levels. The slight increase from the R-squared of the initial model was most likely due to the inclusion of interaction terms. For a better R-squared value, we recommend adding different variables and interaction terms to the linear regression to improve the explained variability of the response by predictors.

Looking at the predictors in the Two-Way Interaction ANOVA Model (Model 1), we also noticed that some races had significant differences in the average blood pressure systolic while others did not. For example, there were a few minority groups that had significant differences in the average systolic blood pressure levels with white people. We expected that the average systolic blood pressure levels of most minority races to differ from that of white people in the model; however, only the average systolic blood pressure levels of black and Mexican-Americans differed significantly from the average systolic blood pressure levels of white people. In order to investigate the presence of these disparities, we recommend redesigning this ANOVA model to include income to poverty, food worries, and insurance coverage as factors in the model. We also looked into the results of this model more by examining the distribution of age

by Race and Income_label and realized that our observations of infants may have skewed our data and recommend creating an ANOVA model without the infants and rerun the post-hoc Tukey HSD tests for the income groups and races in the model.

REFERENCES

Adler, Nancy E, & Newman, Katherine. (2002). Socioeconomic Disparities In Health: Pathways And Policies. Health Affairs, 21(2), 60–76. https://doi.org/10.1377/hlthaff.21.2.60

Centers for Disease Control and Prevention. (2017, January 26). *National Health and Nutrition Examination Survey*. Kaggle. https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey?select=labs.csv.

Devore, J. L. (2018). *Probability and statistics for engineering and the sciences* (8th ed.). Nelson.

Harvard Health. (2021, March 15). *7 ways to reduce stress and keep blood pressure down*. Harvard Health. https://www.health.harvard.edu/heart-health/7-ways-to-reduce-stress-and-keep-blood-pressure-down.

Lackland D. T. (2014). Racial differences in hypertension: implications for high blood pressure management. *The American journal of the medical sciences*, *348*(2), 135–138. https://doi.org/10.1097/MAJ.0000000000000308

Moradi S, Mirzababaei A, Mohammadi H, Moosavian SP, Arab A, Jannat B, Mirzaei K. Food insecurity and the risk of undernutrition complications among children and adolescents: A systematic review and meta-analysis. Nutrition. 2019 Jun;62:52-60. doi: 10.1016/j.nut.2018.11.029. Epub 2018 Dec 7. PMID: 30852458.