Statistical Analysis Project Part II

What factors relate to high blood pressure?

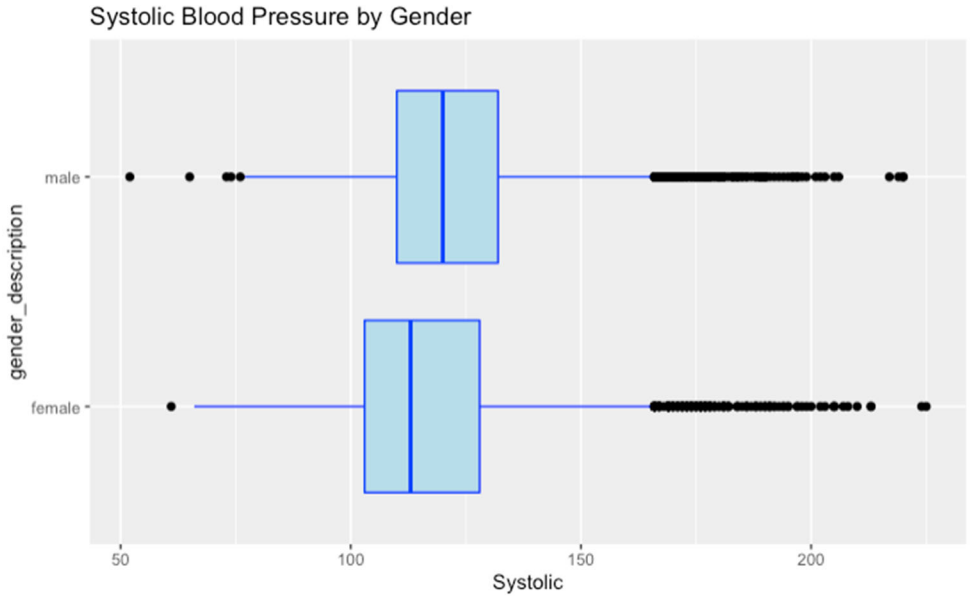
Purpose of Exploratory Analysis

According to Centers for Disease Control and Prevention, nearly half of adults in the United States have hypertension, defined as a systolic blood pressure greater than 130 mmHg or a diastolic blood pressure greater than 80 mmHg. High blood pressure can lead to complications such as stroke, heart attack, and dementia. In order to prevent hypertension and its associated maladies, the data analysis project explores factors that connect to high systolic and diastolic blood pressures.

In Part I of the project, National Health and Nutrition Examination Survey 2017 to March 2020 Pre-Pandemic data was downloaded from Centers for Disease Control and Prevention's website. Potential relationships between blood pressures and factors such as gender, race, education, alcohol consumption, sleep and so on were examined using histograms, box plots, scatter plots, tables, and summary statistics. In this part of the project, Two Sample T-tests will be done to compare blood pressure means of different genders and insurance coverage types. Chi-Square tests will be carried out to compare high blood pressure proportions in different races and in different education levels. Lastly, linear regression analyses will be used to evaluate whether there are linear relationships between blood pressures and income, and between blood pressures and sleep.

Two Sample T-test: Blood Pressures vs Gender

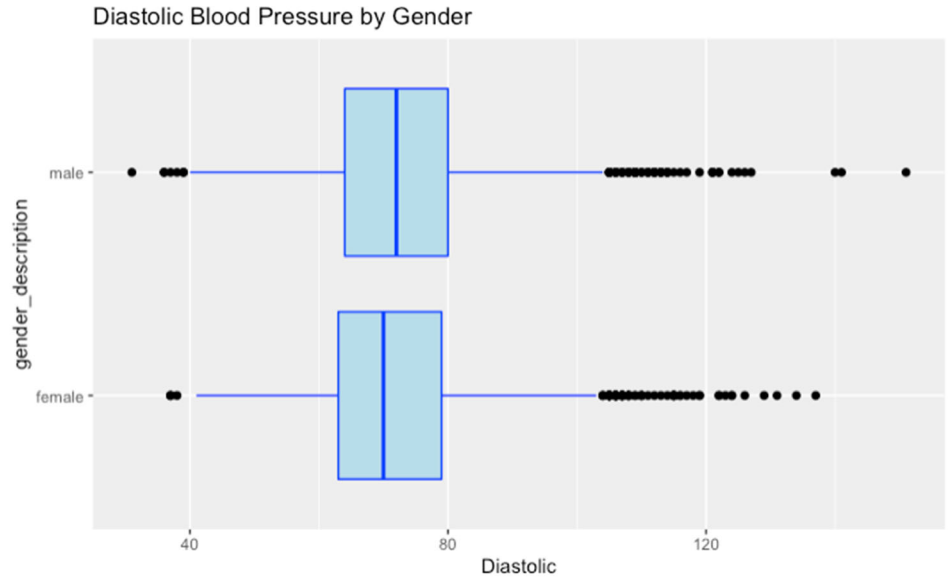
The first tests carried out are Two Sample t-tests. The variables of interest are the two types of blood pressures. The average blood pressures for male and female will be compared in two t-tests - one for systolic, and one for diastolic. The goal of the tests is to determine whether there is statistical evidence in CDC's data that the population mean blood pressures are different for men and women.




```
        Welch Two Sample t-test

data:  Systolic by gender_description
t = -12.506, df = 10270, p-value < 2.2e-16
alternative hypothesis: true difference in means between group female and group male is not equal
to 0
95 percent confidence interval:
 -5.627164 -4.102184
sample estimates:
mean in group female   mean in group male
            117.5892             122.4539
```

For systolic blood pressure, the null hypothesis is that the population systolic blood pressure mean difference between men and women is equal to zero. The two-sided alternative hypothesis is that the population mean difference is not equal to zero. The sample size of female human subjects is 5,932. The sample systolic blood pressure mean for them is 117.59 mmHg with a standard deviation of 20.78 mmHg. The sample size of male participants is 5,724. The sample systolic blood pressure mean for them is 122.45 mmHg with a standard deviation of 18.76 mmHg. The results of the Two Sample t-test show that the t statistic is -12.506 and the corresponding p-value is 2.2e-16. The p-value suggests that given the null hypothesis is true, the

probability of obtaining a sample mean difference as big as 4.86 mmHg and bigger in either direction is very low, much lower than 5%. Because of this, the null hypothesis is rejected in favor of the alternative. There is strong evidence that the population mean difference in systolic blood pressure is not equal to zero. More specifically, the data indicates that the average systolic blood pressure is lower for women than for men. The 95% confidence interval of [-5.63, -4.10] suggests that this range contains the plausible values for the population mean difference between the two genders' systolic blood pressure.



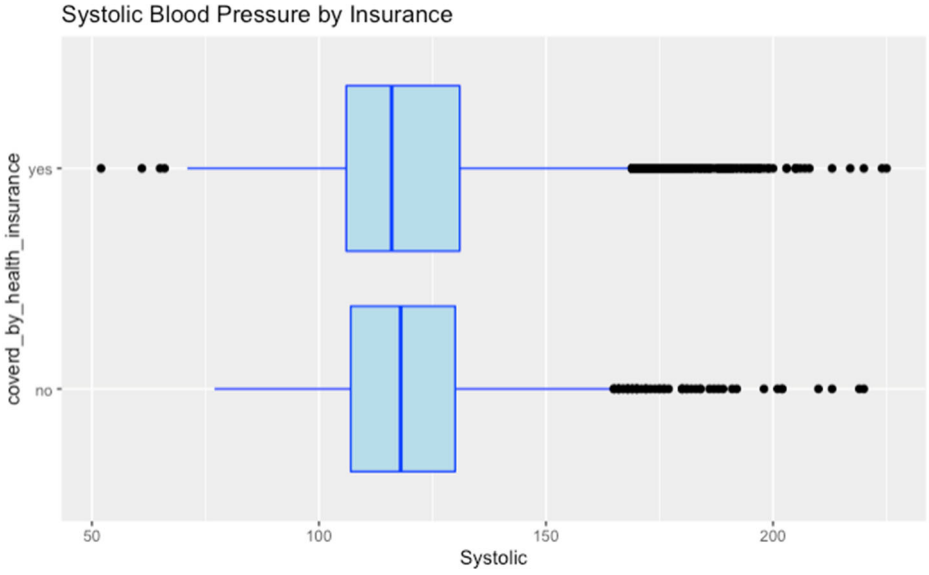

```
        Welch Two Sample t-test

data:  Diastolic by gender_description
t = -3.6193, df = 10332, p-value = 0.0002968
alternative hypothesis: true difference in means between group female and group male is not equal
to 0
95 percent confidence interval:
 -1.3613409 -0.4048084
sample estimates:
mean in group female   mean in group male
            71.59728             72.48035
```

As for diastolic blood pressure, the null hypothesis is that the population diastolic blood pressure mean difference between men and women is equal to zero. The two-sided alternative hypothesis is that the population mean difference is not equal to zero. The sample mean for women is 71.60 mmHg with a standard deviation of 12.24 mmHg. The sample mean for the male counterparts is 72.48 mmHg with a standard deviation of 12.58 mmHg. The results of the Two Sample t-test show that the t statistic is -3.6193 and the corresponding p-value is 0.0002968. The p-value suggests

that given the null hypothesis is true, the probability of obtaining a sample mean difference as big as 0.88 mmHg and bigger in either direction is 0.02968%, which is much smaller than 5%. Because of this, the null hypothesis is rejected in favor of the alternative. There is strong evidence that the population mean difference in diastolic blood pressure is not equal to zero. More specifically, the data indicates that the average diastolic blood pressure is lower for women than for men. The 95% confidence interval suggests that the range of [-1.361, -0.40] contains the plausible values for the population mean difference between the two genders' diastolic blood pressure.

Two Sample T-test: Blood Pressures vs Health Insurance Coverage

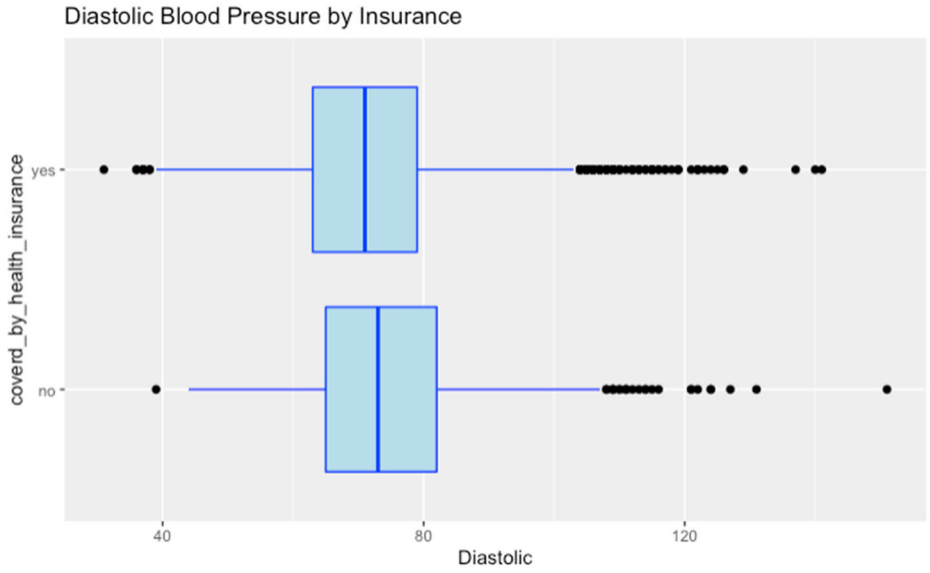
The next tests are Two Sample t-tests where the variables of interest are again the two types of blood pressures. The average blood pressures for participants who have insurance and for those who do not will be compared in two t-tests - one for systolic, and one for diastolic. The goal of the tests is to determine whether there is statistical evidence in the data that the population mean blood pressures are different for those who are covered by insurance and those who are not.




```
        Welch Two Sample t-test

data:  Systolic by coverd_by_health_insurance
t = 2.1272, df = 1842.2, p-value = 0.03353
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 0.094783 2.334828
sample estimates:
 mean in group no mean in group yes
         121.0595          119.8447
```

For systolic blood pressure, the null hypothesis is that the population systolic blood pressure mean difference between people with insurance coverage and people without coverage is equal to zero. The two-sided alternative hypothesis is that the population mean difference is not equal to zero. The sample size for human subjects who answered “yes” to having insurance coverage is 10,031. The sample systolic blood pressure mean for them is 119.84 mmHg with a standard deviation of 19.98 mmHg. The sample size for “no” participants is 1,598. The sample systolic blood

pressure mean for them is 121.06 mmHg with a standard deviation of 19.70 mmHg. The results of the Two Sample t-test show that the t statistic is 2.1272 and the corresponding p-value is 0.03353. The p-value suggests that given the null hypothesis is true, the probability of obtaining a sample mean difference as big as 1.21 mmHg and bigger in either direction is low. Because the p-value is lower than 5%, the null hypothesis is rejected in favor of the alternative. There is significant evidence that the population mean difference in systolic blood pressure is not equal to zero. More specifically, the data indicates that the average systolic blood pressure is lower for people who have insurance coverage. The 95% confidence interval of [0.09, 2.33] suggests that this range contains the plausible values for the population mean difference between the two groups' systolic blood pressure.




```
        Welch Two Sample t-test

data:  Diastolic by coverd_by_health_insurance
t = 7.5038, df = 1760.3, p-value = 9.789e-14
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 2.101401 3.588650
sample estimates:
 mean in group no mean in group yes
         74.49746          71.65244
```

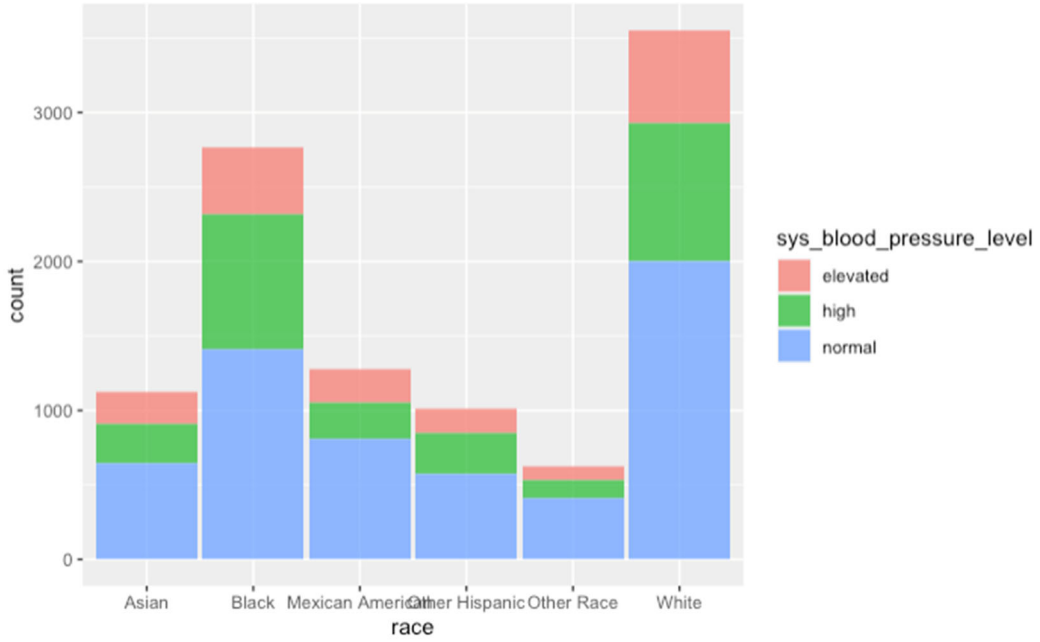
As for diastolic blood pressure, the null hypothesis is that the population diastolic blood pressure mean difference between people with insurance coverage and people without coverage is equal to zero. The two-sided alternative hypothesis is that the population mean difference is not equal to zero. The sample mean for participants who answered "yes" to having insurance coverage is 71.65 mmHg with a standard

deviation of 12.23 mmHg. The sample mean for the “no” counterparts is 74.50 mmHg with a standard deviation of 13.23 mmHg. The results of the Two Sample t-test show that the t statistic is 7.50 and the corresponding p-value is 9.79e-14. The p-value suggests that given the null hypothesis is true, the probability of obtaining a sample mean difference as big as 2.85 mmHg and bigger in either direction is extremely low. Because of this, the null hypothesis is rejected in favor of the alternative. There is strong evidence that the population mean difference in diastolic blood pressure is not equal to zero. More specifically, the data indicates that the average diastolic blood pressure is significantly lower for people who have insurance coverage than people who do not. The 95% confidence interval suggests that the range of [2.10, 3.59] contains the plausible values for the population mean difference between the two groups’ diastolic blood pressure.

Chi-Square Test: Blood Pressures vs. Race

Every interviewee's systolic and diastolic blood pressures are categorized into a blood pressure level based on the American College of Cardiology/American Heart Association Guideline published in 2017. A systolic blood pressure can fall into normal, elevated, or high level. And a diastolic blood pressure can be categorized as either normal or high.

For the first two Chi-Square tests, the variables of interest are high systolic and diastolic blood pressures. The tests compare each race's high blood pressure proportions to see if there is statistical evidence that indicates an association between high blood pressures and race.



Observed Counts

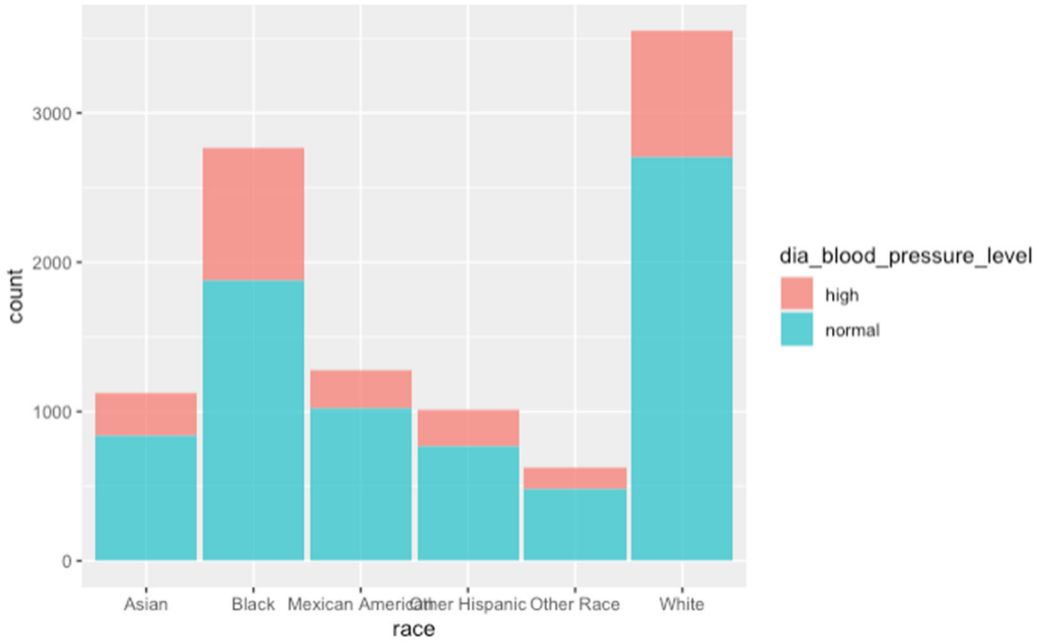
	Asian	Black	Mexican American	Other Hispanic	Other Race	White
elevated	212	449	222	168	94	620
high	267	910	238	269	120	929
normal	640	1408	812	575	414	2005

Expected Counts

race$sys_blood_pressure_level	race$race_description Asian	Black	Mexican American	Other Hispanic	Other Race	White
elevated	190.7878	471.7692	216.8740	172.5444	107.0730	605.9515
high	295.4238	730.5072	335.8168	267.1750	165.7964	938.2807
normal	632.7884	1564.7235	719.3091	572.2805	355.1306	2009.7678

```
        Pearson's Chi-squared test

data:  race$sys_blood_pressure_level and race$race_description
X-squared = 131.21, df = 10, p-value < 2.2e-16
```

For systolic blood pressure, the null hypothesis is there is no association between high systolic blood pressure and race. The alternative is there is an association between high systolic blood pressure and race. The result of the Chi-Square Test produced a p-value of 2.2e-16, meaning given the null hypothesis is true, the chance of observing a sample with such count distribution as presented by the sample (see the Observed Counts table) is extremely small. Therefore, the null hypothesis is rejected in favor of the alternative. There is strong statistical evidence that high blood pressure and race are not independent of each other. More specifically, the count distribution presented in the Expected Counts table suggests that the proportion in the population that suffer high systolic blood pressure is greater for Black than for the rest of the races.



Observed Counts

	Asian	Black	Mexican American	Other Hispanic	Other Race	White
high	276	891	247	242	145	848
normal	843	1876	1025	770	483	2706

Expected Counts

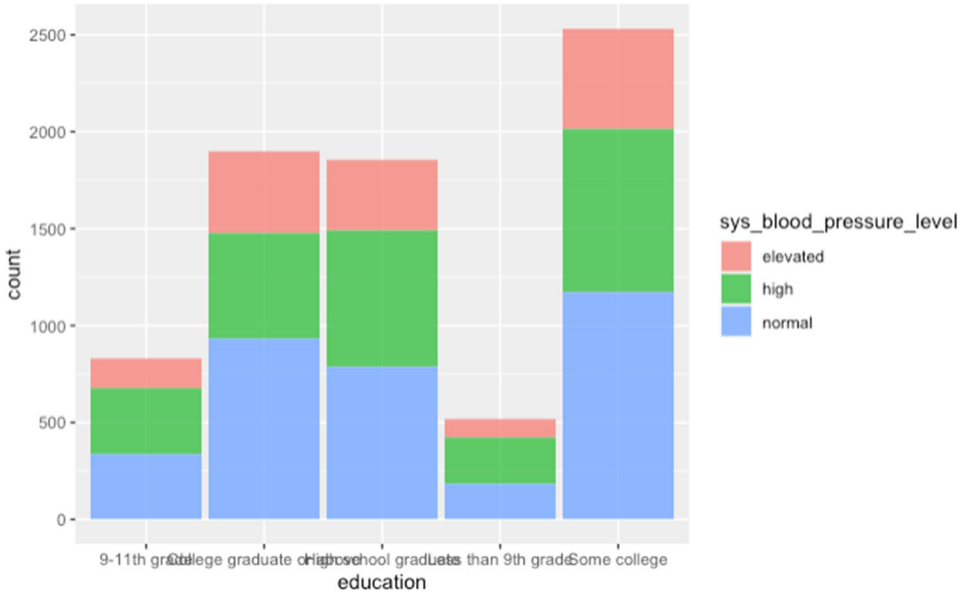
	race$race_description					
race$dia_blood_pressure_level	Asian	Black	Mexican American	Other Hispanic	Other Race	White
high	286.3438	708.0548	325.4954	258.9633	160.7005	909.4422
normal	832.6562	2058.9452	946.5046	753.0367	467.2995	2644.5578

```
        Pearson's Chi-squared test

data:  race$dia_blood_pressure_level and race$race_description
X-squared = 98.599, df = 5, p-value < 2.2e-16
```

For diastolic blood pressure, the null hypothesis is there is no association between high diastolic blood pressure and race. The alternative is there is an association between high diastolic blood pressure and race. The result of the Chi-Square Test produced a p-value of 2.2e-16, meaning given the null hypothesis is true, the chance of observing a sample with such count distribution as presented by the data (see the Observed Counts table) is extremely small. Therefore, the null hypothesis is rejected in favor of the alternative. There is strong statistical evidence of an association between high blood pressure and race. More specifically, the count distribution presented in the Expected Counts table shows the proportion in the population that suffer high diastolic blood pressure is greater for Black than for the rest of the races.

Chi-Square Test: Blood Pressures vs. Education

In the next two Chi-Square tests, the variables of interest are again high systolic and diastolic blood pressures. The tests compare the proportions of high blood pressures of different education levels to see if there is statistical evidence that indicates an association between high blood pressures and education.



Observed Counts

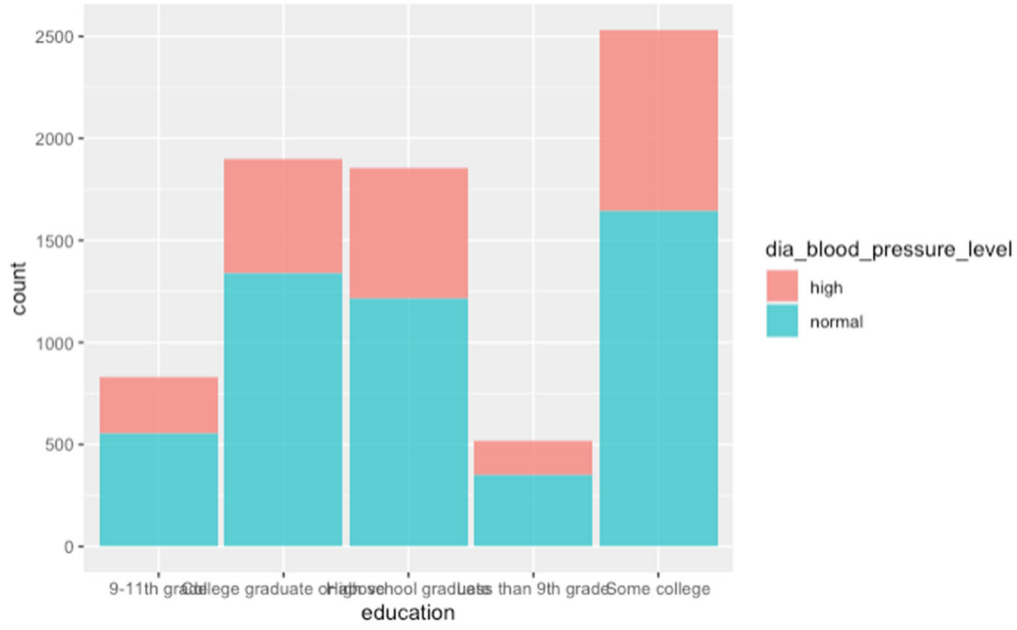
	9-11th grade	College graduate or above	High school graduate	Less than 9th grade	Some college
elevated	152	423	366	92	519
high	343	548	702	245	841
normal	338	930	790	181	1173

Expected Counts

	education$education_level				
	9-11th grade	College graduate or above	High school graduate	Less than 9th grade	Some college
elevated	169.1503	386.0201	377.2885	105.1859	514.3551
high	291.9805	666.3325	651.2602	181.5677	887.8591
normal	371.8692	848.6474	829.4513	231.2464	1130.7858

```
        Pearson's Chi-squared test

data:  education$sys_blood_pressure_level and education$education_level
X-squared = 91.084, df = 8, p-value = 2.802e-16
```

For systolic blood pressure, the null hypothesis is there is no association between high systolic blood pressure and education. The alternative is there is an association between high systolic blood pressure and education. The results of the Chi-Square Test show a p-value of 2.802e-16, meaning given the null hypothesis is true, the chance of observing a sample with such count distribution as presented by the data (see the Observed Counts table) is extremely small. Therefore, the null hypothesis is rejected in favor of the alternative. There is strong statistical evidence that high blood pressure and education are not independent of each other. More specifically, compared to the count distribution presented in the Expected Counts table, the proportion in the population that suffer high systolic blood pressure is greater for people with high school and below education than people in the other education groups.



Observed Counts

	9-11th grade	College graduate or above	High school graduate	Less than 9th grade	Some college
high	279	564	641	168	892
normal	554	1337	1217	350	1641

Expected Counts

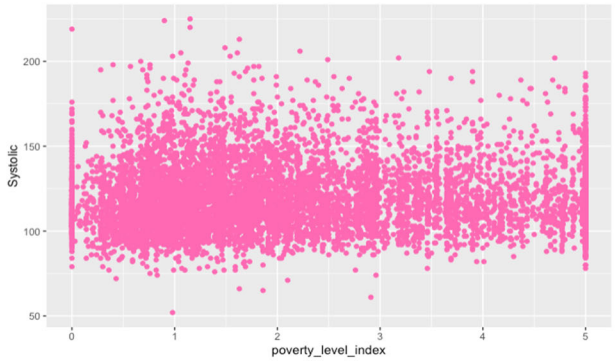
	education$education_level				
	9-11th grade	College graduate or above	High school graduate	Less than 9th grade	Some college
high	277.267	632.7547	618.442	172.4182	843.1181
normal	555.733	1268.2453	1239.558	345.5818	1689.8819

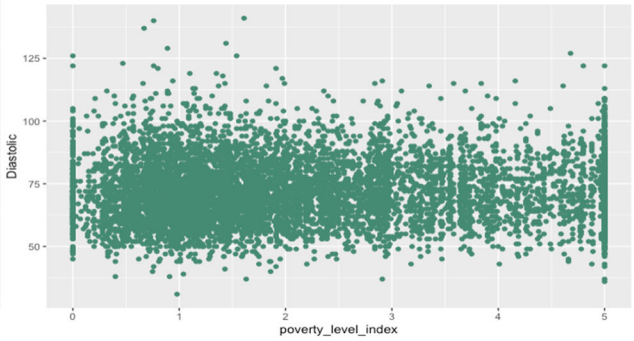
```
        Pearson's Chi-squared test

data:  education$dia_blood_pressure_level and education$education_level
X-squared = 16.865, df = 4, p-value = 0.002053
```

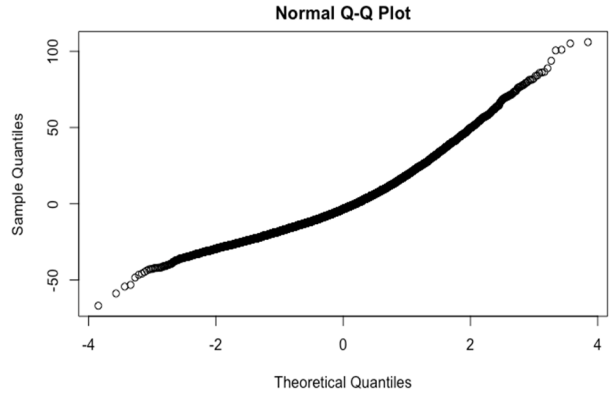
For diastolic blood pressure, the null hypothesis is there is no association between high diastolic blood pressure and education. The alternative is there is an association between high diastolic blood pressure and education. The results of the Chi-Square Test indicate a p-value of 0.002053, meaning given the null hypothesis is true, the chance of observing a sample with such count distribution as presented by the data (see the Observed Counts table) is 0.2053%, which is smaller than 5%. Therefore, the null hypothesis is rejected in favor of the alternative. There is strong statistical evidence of an association between high blood pressure and education level. More specifically, compared to the count distribution presented in the Expected Counts table, the proportion in the population that suffer high diastolic blood pressure is greater for people with 9th-grade to some-college level of education than people in the other education groups.

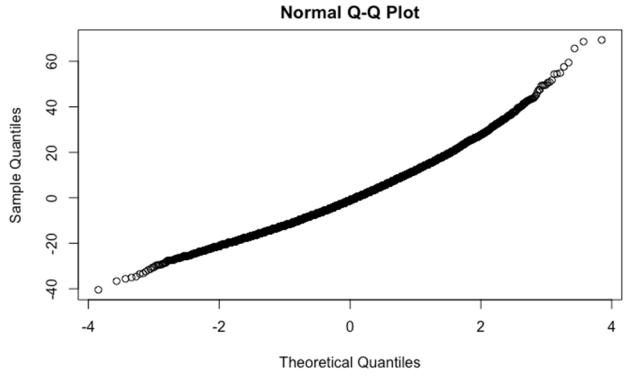
Linear Regression Analysis: Blood Pressures vs. Income

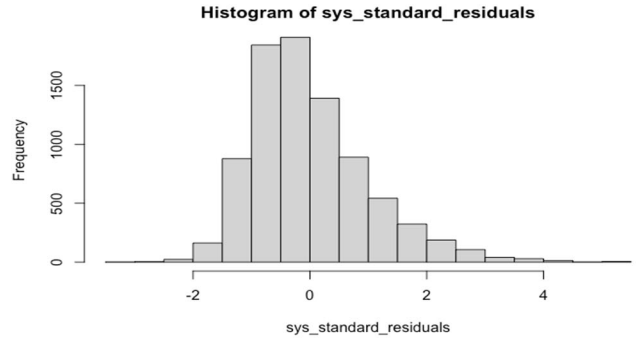
To explore the relationship between blood pressures and income, linear regression analyses are done where poverty level index is the explanatory variable and blood pressures are the response variables. The correlation between systolic blood pressure and poverty level index is 0.038, which is weak. The correlation between diastolic blood pressure and poverty level index is also weak - 0.037.

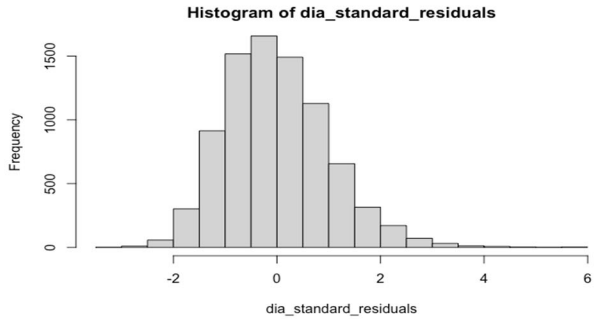


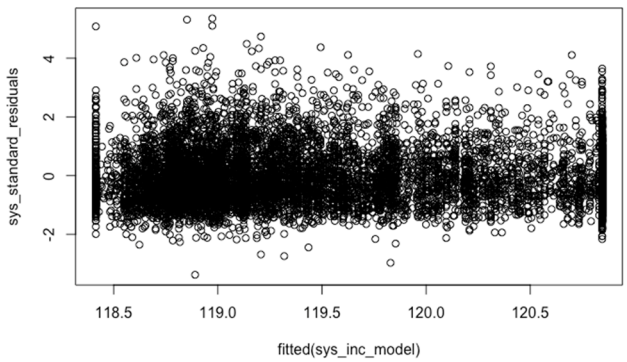


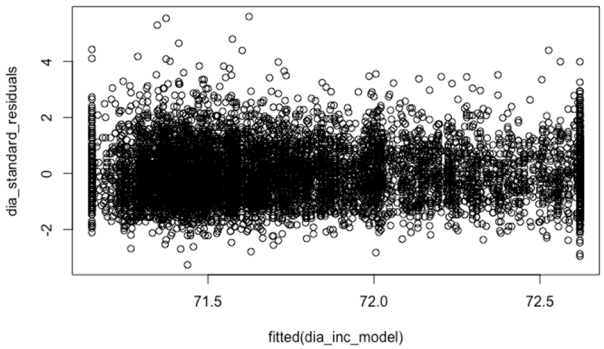




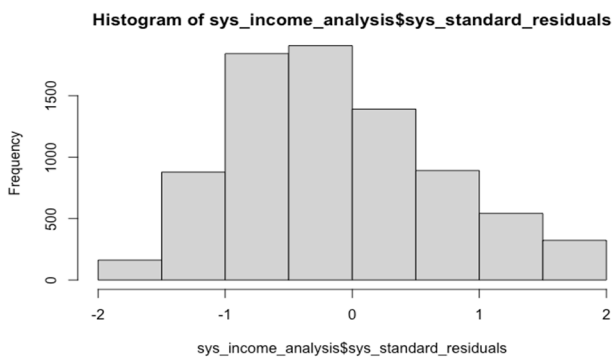


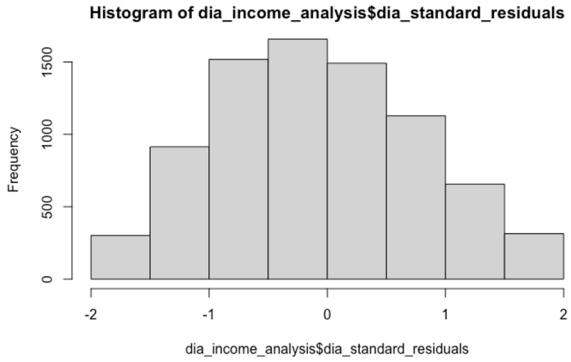




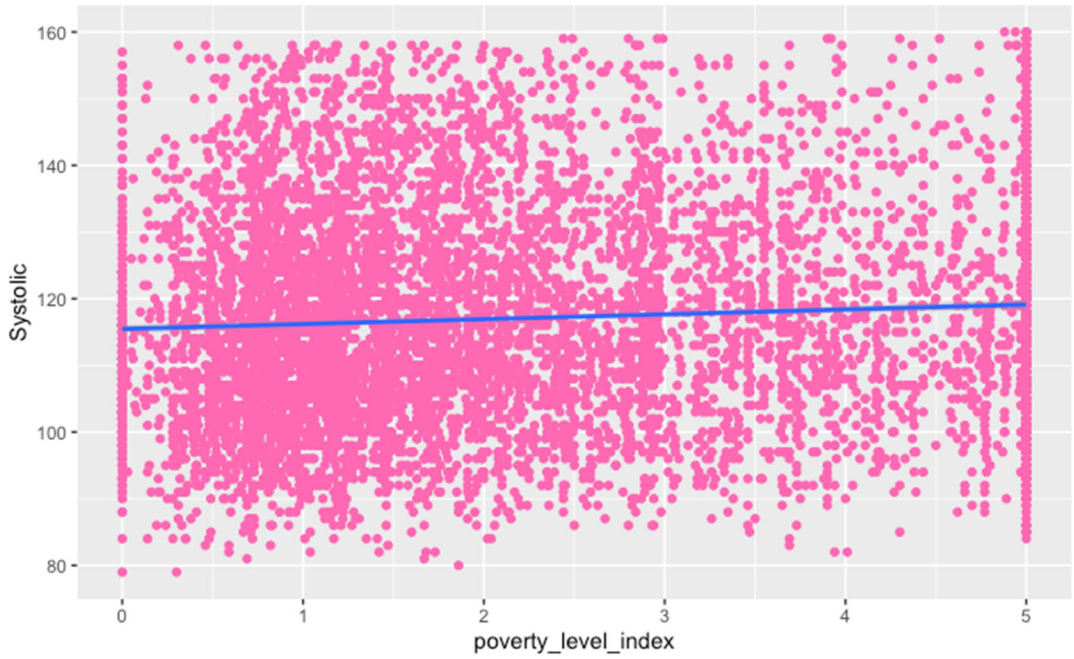


CDC's data meets linear regression analysis' assumption requirements. First, based on the scatter plots, both trends, systolic blood pressure vs. poverty level index and diastolic blood pressure vs. poverty level index, appear to be straight. Second, the data is random because the survey conductor intentionally surveyed from a randomized, nationally representative sample. Third, although both of the scatter plots appear to be thicker around the lower numbers on the x-axis, manipulation will be done to remove outliers that are bigger than 2 standardized residuals to attain as much common variance as possible. Lastly, after the removal outliers, the histograms of error terms show normal distributions.



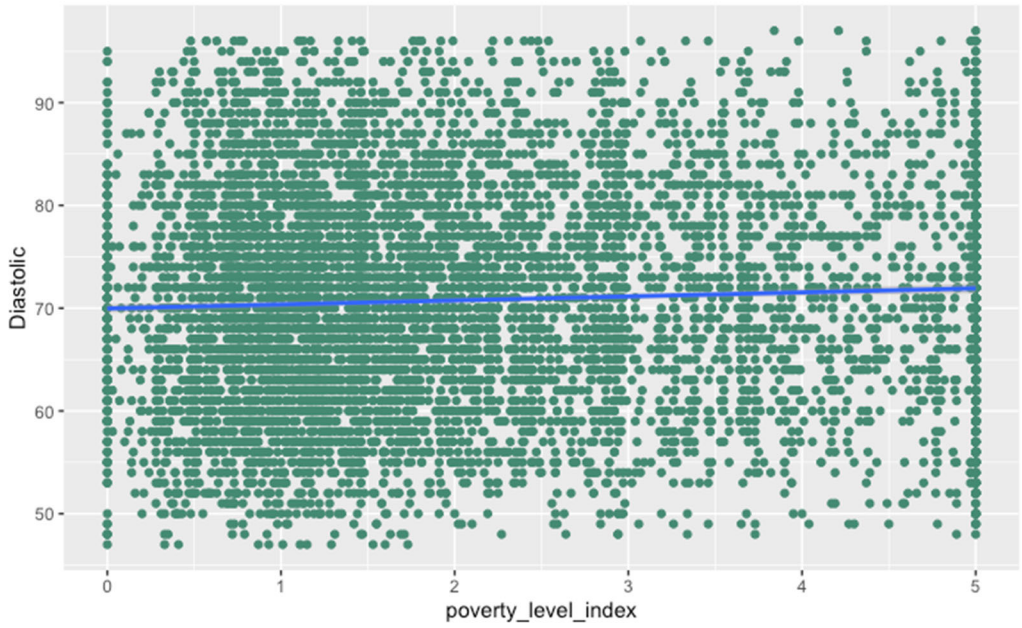


The null hypotheses of the tests are that the population slopes are equal to zero. In other words, there is no linear relationship between systolic, diastolic blood pressures and poverty level index. The alternative hypotheses are the population slopes are not equal to zero; there are linear relationships between the two types of blood pressures and poverty level index.




```
Call:
lm(formula = Systolic ~ poverty_level_index, data = sys_income_analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-36.823 -11.863  -2.114  10.445  42.308

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        115.4656     0.3200 360.848  < 2e-16 ***
poverty_level_index   0.7297     0.1153   6.331 2.56e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.02 on 7936 degrees of freedom
Multiple R-squared:  0.005026,  Adjusted R-squared:  0.004901
F-statistic: 40.09 on 1 and 7936 DF,  p-value: 2.561e-10
```

The results of the analysis for systolic blood pressure and poverty level index show that the estimate for intercept is 115.47 mmHg. The t statistic is 360.85, which gives a p-value of 2e-16. The slope estimate is 0.73 mmHg/unit of poverty-level-index. The t statistic is 6.33, which gives a p-value of 2.56e-10. The low p-value indicates that there is strong statistical evidence in the data of a linear relationship between systolic blood pressure and poverty level index. For every one unit increase in the poverty level index, one’s systolic blood pressure increases by 0.73 mmHg. Although this direct, linear relationship suggests a connection between high income and high systolic blood pressure, the R-Square of the test is merely 0.005026, meaning only 0.5026% of the variation in systolic blood pressure can be explained by a linear relationship with poverty index level.

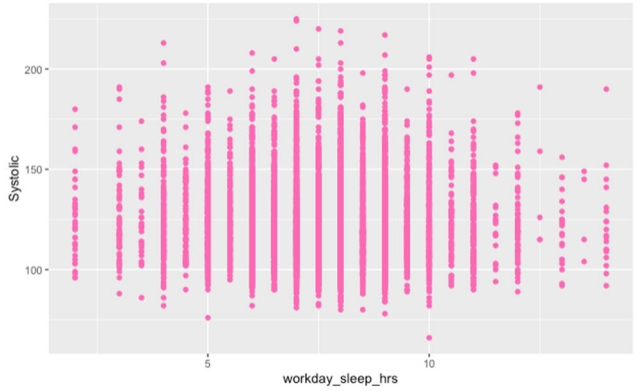


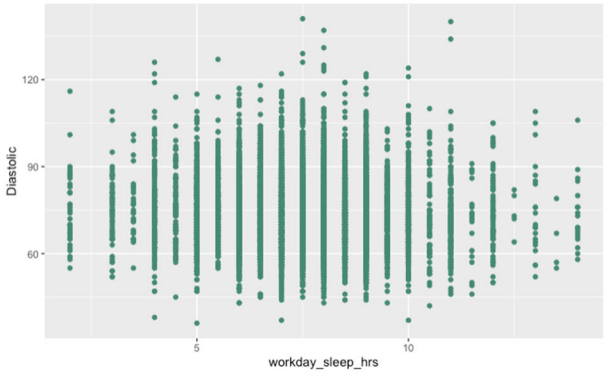

```
Call:
lm(formula = Diastolic ~ poverty_level_index, data = dia_income_analysis)

Residuals:
     Min       1Q   Median       3Q      Max
-23.9257  -7.9257  -0.5485   7.5868  25.8428

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        69.96940    0.21140 330.978  < 2e-16 ***
poverty_level_index  0.39127    0.07652   5.113 3.24e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.6 on 7979 degrees of freedom
Multiple R-squared:  0.003266,  Adjusted R-squared:  0.003141
F-statistic: 26.15 on 1 and 7979 DF,  p-value: 3.236e-07
```

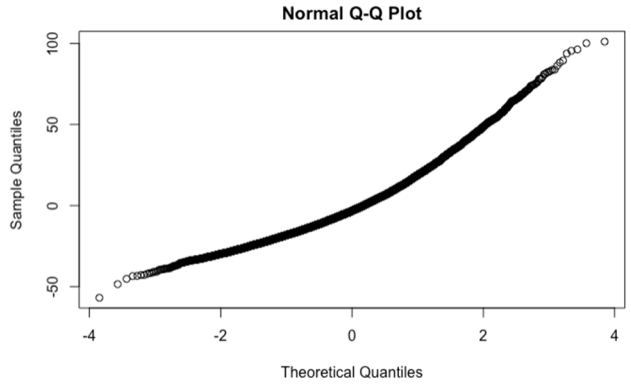
As for the analysis of diastolic blood pressure and poverty level index, the results show that the estimate for intercept is 69.97 mmHg. The t statistic is 330.98, which gives a p-value of 2e-16. The slope estimate is 0.39 mmHg/unit of poverty-level-index. The t statistic is 5.11, which gives a p-value of 3.24e-07. The low p-value indicates that there is strong statistical evidence in the data of a linear relationship between diastolic blood pressure and poverty level index. For every one unit increase in the poverty level index, one’s diastolic blood pressure increases by 0.39 mmHg. Although this direct relationship suggests a connection between high income and high diastolic blood pressure, the R-Square of the test is merely 0.003266, meaning only 0.3266% of the variation in diastolic blood pressure can be explained by a linear relationship with poverty index level.

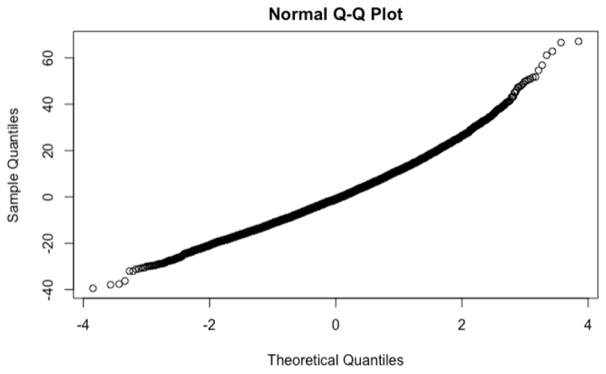
Linear Regression Analysis: Blood Pressures vs. Sleep

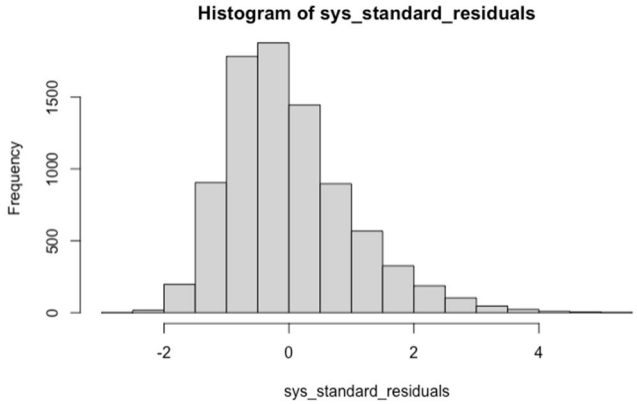
To explore the relationships between blood pressures and sleep, linear regression analyses are done where the number of workday sleep hours is the explanatory variable and blood pressures are the response variables. The correlation between systolic blood pressure and workday sleep hours is -0.027, which is weak. The correlation between diastolic blood pressure and poverty level index is also weak - -0.061.

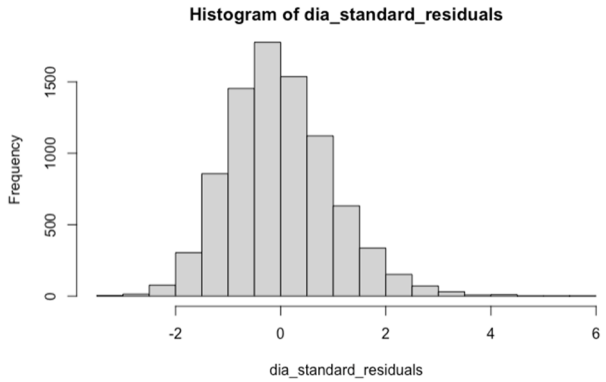


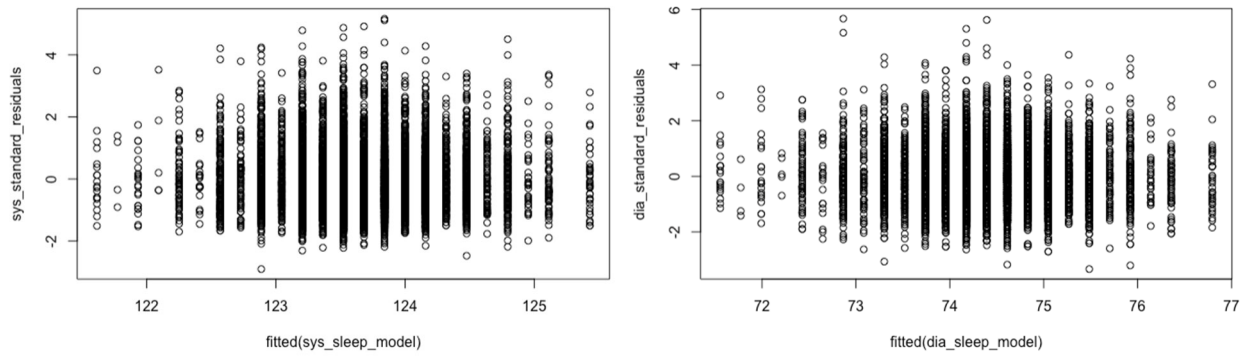




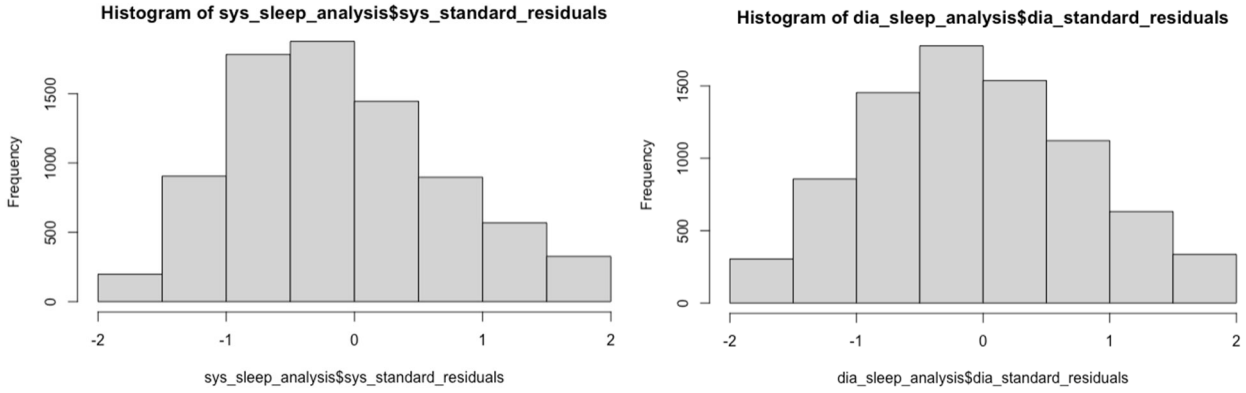




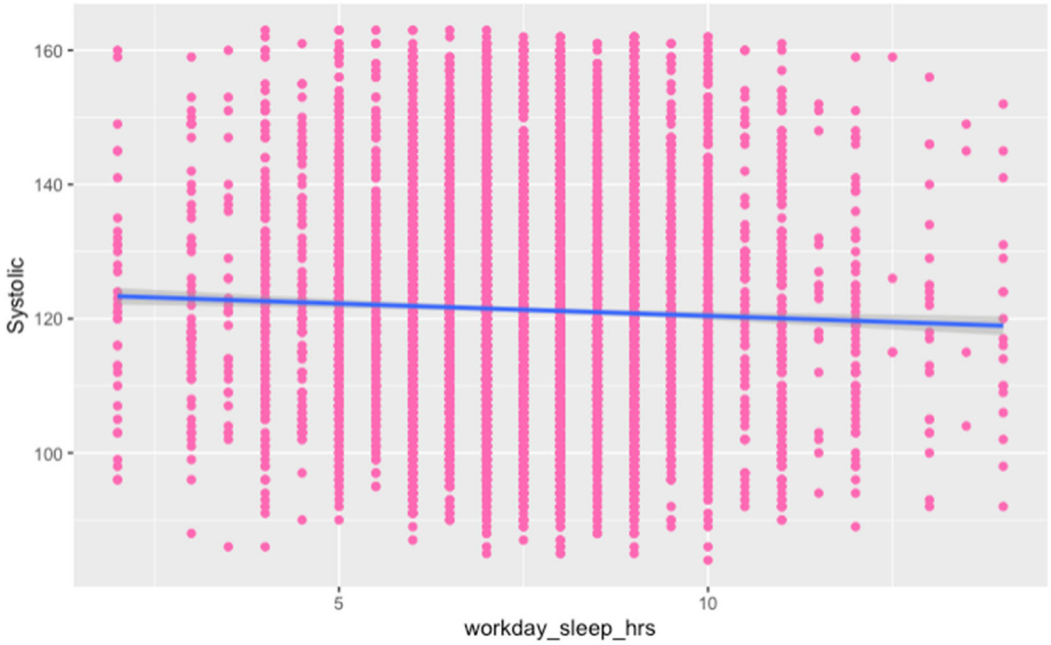




CDC's data meets linear regression analysis' assumption requirements. First, based on the scatter plots, both trends, systolic blood pressure vs. workday sleep hours and diastolic blood pressure vs. workday sleep hours, appear to be straight. Second, the data is random because the survey conductor intentionally surveyed from a randomized, nationally representative sample. Third, although both of the scatter plots appear to be thicker around 6-8 hours (of sleep) on the x-axis, manipulation will be done to remove outliers that are over 2 standardized residuals to achieve as much common variance as possible. Lastly, after the removal outliers, the histograms of error terms show normal distributions.

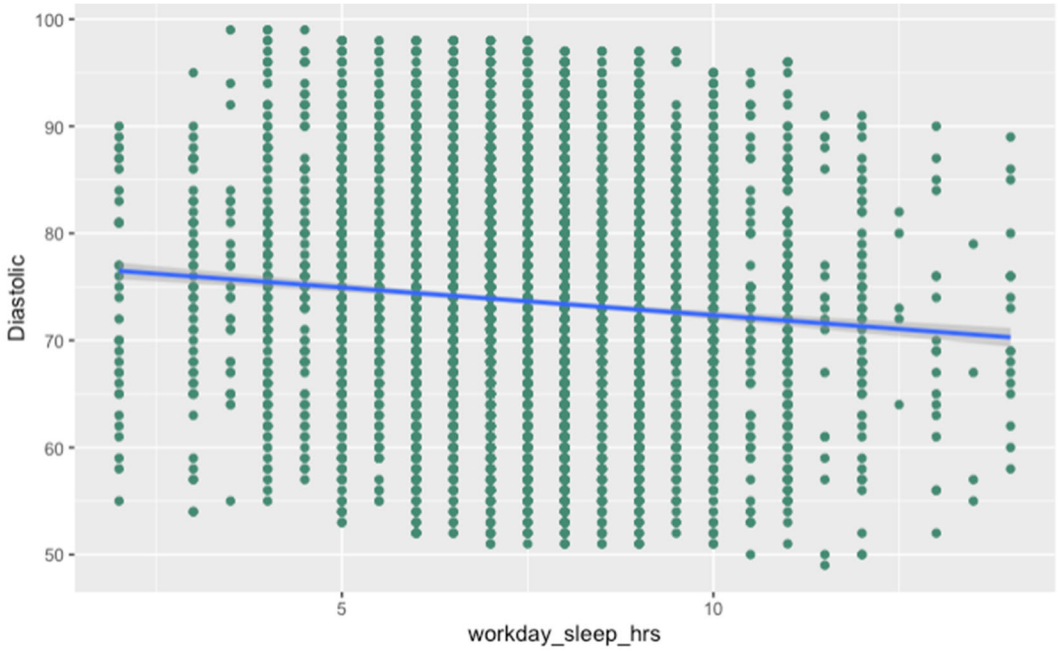


The null hypotheses of the tests are that the population slopes are equal to zero. In other words, there is no linear relationship between systolic, diastolic blood pressures and workday sleep hours. The alternative hypotheses are the population slopes are not equal to zero; there are linear relationships between the two types of blood pressures and workday sleep hours.




```
Call:
lm(formula = Systolic ~ workday_sleep_hrs, data = sys_sleep_analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-36.793 -11.965  -1.879  10.465  41.584

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       124.0730     0.8408 147.561  < 2e-16 ***
workday_sleep_hrs  -0.3657     0.1081  -3.381 0.000725 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.06 on 7996 degrees of freedom
Multiple R-squared:  0.001428,  Adjusted R-squared:  0.001303
F-statistic: 11.43 on 1 and 7996 DF,  p-value: 0.0007251
```

The results of the analysis for systolic blood pressure and workday sleep hours show that the estimate for intercept is 124.07 mmHg. The t statistic is 147.56, which gives a p-value of 2e-16. The slope estimate is -0.37 mmHg/hour of sleep. The t statistic is -3.381, which gives a p-value of 0.000725. The low p-value indicates that there is strong statistical evidence in the data of a linear relationship between systolic blood pressure and sleep. For every additional hour of workday sleep, one’s systolic blood pressure decreases by 0.37 mmHg. Although this inverse, linear relationship suggests a connection between little sleep and high systolic blood pressure, the R-Square of the test is merely 0.001428, meaning only 0.1428% of the variation in systolic blood pressure can be explained by a linear relationship with workday sleep hours.




```
Call:
lm(formula = Diastolic ~ workday_sleep_hrs, data = dia_sleep_analysis)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9057  -7.4218  -0.4218   7.0943  24.3847

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       77.51872    0.52905 146.524  < 2e-16 ***
workday_sleep_hrs -0.51615    0.06807  -7.583 3.76e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.08 on 8017 degrees of freedom
Multiple R-squared:  0.007121,  Adjusted R-squared:  0.006997
F-statistic:  57.5 on 1 and 8017 DF,  p-value: 3.76e-14
```

As for the analysis of diastolic blood pressure and workday sleep hours, the results show that the estimate for intercept is 77.52 mmHg. The t statistic is 146.52, which gives a p-value of 2e-16. The slope estimate is -0.52 mmHg/hour of sleep. The t statistic is -7.58, which gives a p-value of 3.76e-14. The low p-value indicates that there is strong statistical evidence in the data of a linear relationship between diastolic blood pressure and workday sleep hours. For every additional hour of workday sleep, one’s diastolic blood pressure decreases by 0.52 mmHg. Although this inverse relationship suggests a connection between little sleep and high diastolic blood pressure, the R-Square of the test is merely 0.007121, meaning only 0.7121% of the variation in diastolic blood pressure can be explained by a linear relationship with workday sleep hours.

Conclusion

Results of the tests demonstrate that male, lack of insurance coverage, Black, high-school-and-below education, high income, and little sleep are all features that relate to high systolic blood pressure, although less than 1% of the variation in systolic blood pressure can be explained by linear relationships with poverty index level and with workday sleep hours. In order to explain more variation, multiple regression analyses may be necessary to explore hidden trends or potential confounders for the relationships between systolic blood pressure and income, and between systolic blood pressure and sleep.

The sample data also provides significant evidence to show that high diastolic blood pressure and the following factors are related: male, lack of insurance coverage, Black, 9th-grade to some-college level of education, high income, and little sleep, although less than 1% of the variation in diastolic blood pressure can be explained by linear relationships with poverty index level and with workday sleep hours. Similar to systolic blood pressure, multiple regression analyses may be helpful in explaining more variation by exploring hidden trends or potential confounders for the relationships between diastolic blood pressure and income, and between diastolic blood pressure and sleep.