Final Project

Group 3: Amina Lampkin, Zainab Sherani, Linette Kingston, Hayden Smedley

Dataset: *Behavioral Risk Factor Surveillance System (BRFSS)*

- Self-administered questionnaire survey
- Data collected through in-depth interviews via phone, mail, email, and in-person
- Reflects the year 2018

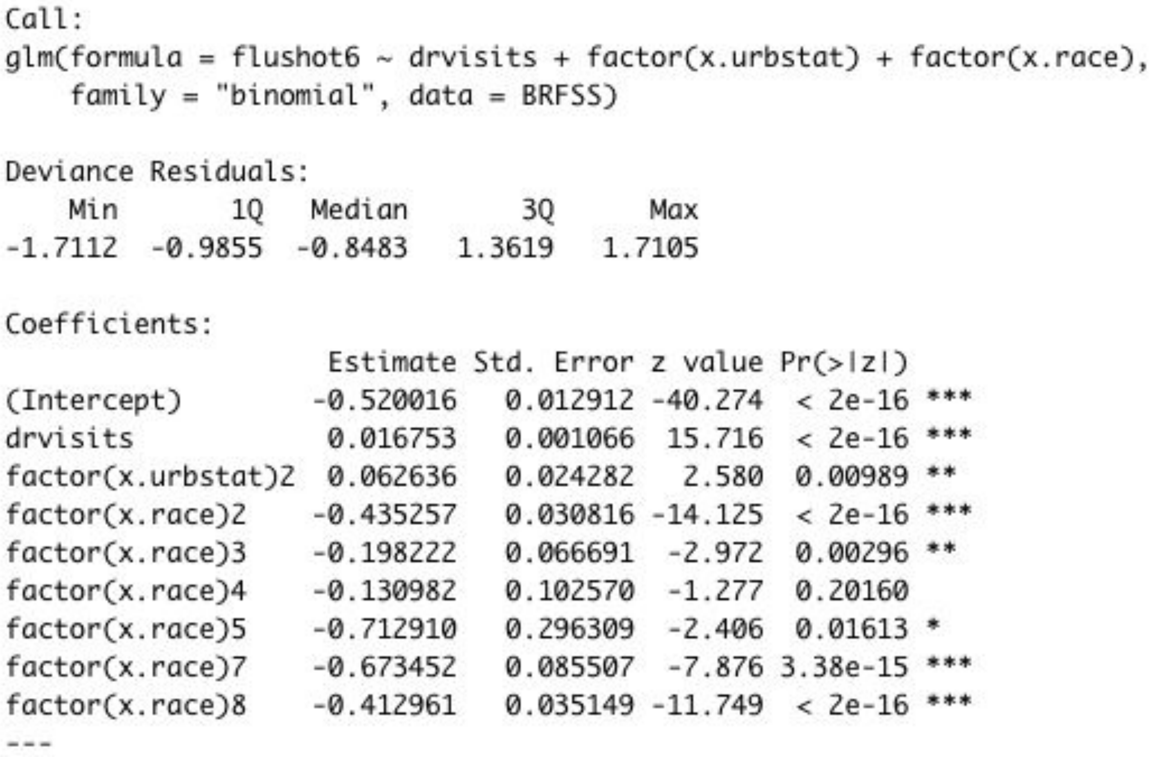
Research Q1: *Does whether or not someone got a flu shot/spray in the past year depend on their residence in an urban or rural county, race, or the number of times they've visited a health professional in the past year?*

- **Independent Variables:**
  - drvisits - # of times in the past year you've seen a health professional
  - x.urbstat - residence in an urban or rural county
  - x.race - race
- **Dependent Variable**
  - flushot6 - whether or not you got a flu shot/spray in the past year

Q1: Data Cleaning

- drvisits recode
  - **88** (visited a health professional zero times) to **0**
  - **77** ("don't know/not sure") to **NA**
- x.race recode
  - **9** (represents "don't know/not sure/refused") and **6** ("other race only, Non-Hispanic") to **NA**
- flushot6 recode
  - **2** (not having gotten the flu shot/spray) to **0**
  - **9** ("refused") and 7 ("Don't know/not sure") to **NA**

Q1: Logistic Regression Results



```
Call:
glm(formula = flushot6 ~ drvisits + factor(x.urbstat) + factor(x.race),
    family = "binomial", data = BRFSS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7112  -0.9855  -0.8483   1.3619   1.7105

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.520016   0.012912 -40.274  < 2e-16 ***
drvisits           0.016753   0.001066  15.716  < 2e-16 ***
factor(x.urbstat)2 0.062636   0.024282   2.580  0.00989 **
factor(x.race)2   -0.435257   0.030816 -14.125  < 2e-16 ***
factor(x.race)3   -0.198222   0.066691  -2.972  0.00296 **
factor(x.race)4   -0.130982   0.102570  -1.277  0.20160
factor(x.race)5   -0.712910   0.296309  -2.406  0.01613 *
factor(x.race)7   -0.673452   0.085507  -7.876 3.38e-15 ***
factor(x.race)8   -0.412961   0.035149 -11.749  < 2e-16 ***
---
```

Q1: Interpreting Slope Coefficients

By number of health visits
(drvisits)

By urban/rural county residence
(x.urbstat)2

- A 1 visit increase in the *number of times an individual has visited a health professional* in the past year causes the log odds of one having received a flu shot/spray in the past year to increase by 0.016753, holding the other independent variables constant.

- The log odds of having gotten a flu shot/spray in the past year between *those who live in a rural county and those who live in an urban county* is 0.062636, holding the other independent variables constant.

Q1: Interpreting Slope Coefficients

By race

(x.race)2

(x.race)3

(x.race)4

- The log odds of having gotten a flu shot/spray in the past year between those who are *Non-Hispanic Black,* and those who are White only, Non-Hispanic is -0.435257, holding the other independent variables constant.
- The log odds of having gotten a flu shot/spray in the past year between those who are *American Indian or Alaskan Native only, Non-Hispanic* and those who are White only, Non-Hispanic is -0.198222, holding the other independent variables constant.
- The log odds of having gotten a flu shot/spray in the past year in the past year between those who are *Non-Hispanic Asian* and those who are White only, Non-Hispanic is -0.130982, holding the other independent variables constant.

Q1: Interpreting Slope Coefficients

By race

(x.race)5

(x.race)7

(x.race)8

- The log odds of having gotten a flu shot/spray in the past year between those who are *Native Hawaiian or Pacific Islander only, Non-Hispanic* and those who are White only, Non-Hispanic is -0.712910, holding the other independent variables constant.
- The log odds of having gotten a flu shot/spray in the past year between those who are *multiracial, Non-Hispanic* and those who are White only, Non-Hispanic is -0.673452, holding the other independent variables constant.
- The log odds of having gotten a flu shot/spray in the past year between those who are *Hispanic* and those who are White only, Non-Hispanic is -0.412961, holding the other independent variables constant.

Q1: Interpreting P-values Overview

- Predictors that WERE significant in predicting whether someone received a flu shot/spray in the past year
    - Number of times visited a health professional in the last year
    - Residence in a rural county
    - Native Hawaiian or other Pacific Islander only, Non-Hispanic
    - Black only, Non-Hispanic
    - American Indian or Alaskan Native only, Non-Hispanic
    - Multiracial, Non-Hispanic
    - Hispanic
- Predictors that were NOT significant
    - Asian only, Non-Hispanic

Q1: Interpreting P-Values

number of health visits
(drvisits)

rural county residence
(x.urbstat)2

- Since the p-value for "the number of times an individual has visited a health professional in the past year" (2e-16) is less than alpha, 0.05, we can conclude that the *number of times an individual has visited a health professional in the past year* is associated with whether or not an individual got a flu shot or spray in the past year.
- Since the p-value for "rural counties" (0.00989) is less than alpha, 0.05, we can conclude that *residence in a rural county* is associated with whether or not an individual got a flu shot or spray in the past year.

Q1: Interpreting P-Values

By race

(x.race)2

(x.race)3

(x.race)4

- Since the p-value for "Black only, Non-Hispanic" (<2e-16) is less than alpha, 0.05, we can conclude that whether someone is *non-Hispanic Black* is associated with whether or not they got a flu shot or spray in the past year.
- Since the p-value for "American Indian or Alaskan Native only, Non-Hispanic" (0.00296) is less than alpha, 0.05, we can conclude that whether someone is *American Indian or Alaskan Native only, Non-Hispanic* is associated with whether or not they got a flu shot or spray in the past year.
- Since the p-value for "Asian only, Non-Hispanic" (0.20160) is less than alpha, 0.05, we can conclude that whether someone is *Non-Hispanic Asian* is NOT associated with whether or not they got a flu shot or spray in the past year.

Q1: Interpreting P-Values

By race

(x.race)5

(x.race)7

(x.race)8

- Since the p-value for "Native Hawaiian or other Pacific Islander only, Non-Hispanic" (0.01613) is less than alpha, 0.05, we can conclude that whether someone is *Native Hawaiian or other Pacific Islander only, Non-Hispanic* is associated with whether or not they got a flu shot or spray in the past year.
- Since the p-value for "Multiracial, Non-Hispanic" (3.38e-15) is less than alpha, 0.05, we can conclude that whether someone is *Multiracial, Non-Hispanic* is associated with whether or not they got a flu shot or spray in the past year.
- Since the p-value for "Hispanic" (<2e-16) is less than alpha, 0.05, we can conclude that whether someone is *Hispanic* is associated with whether or not they got a flu shot or spray in the past year.

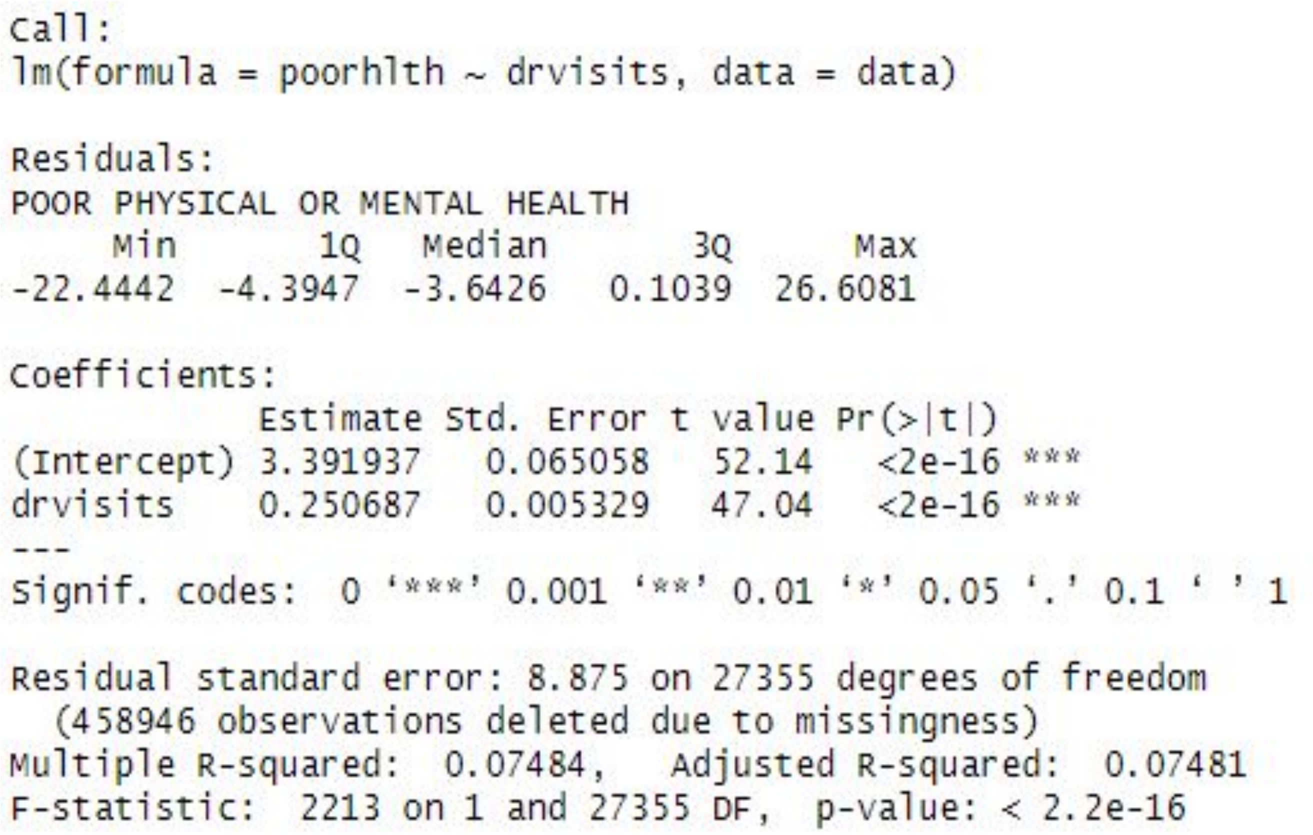
Research Q2: *How does how a person rate their health differ based on the frequency of health checkups?*

- **Independent variable:**
  - drvisits - # of times in the past year you've seen a health professional
- **Dependent variable:**
  - poorhlth - # of days in the last 30 days that are rated of poor mental or physical health

Q2: Data Cleaning

- drvisits recode
  - Recode 77 (Don’t Know/Not Sure) to NA
  - Recode 88 (None) to 0
- poorhlth recode
  - Recode 88 (None) to 0
  - Recode 77 (Don’t Know/Not Sure) to NA
  - Recode 99 (Refused) to NA

Q2: Simple Linear Regression Results



```
Call:
lm(formula = poorhlth ~ drvisits, data = data)

Residuals:
POOR PHYSICAL OR MENTAL HEALTH
     Min       1Q   Median       3Q      Max 
-22.4442  -4.3947  -3.6426   0.1039  26.6081 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.391937   0.065058   52.14   <2e-16 ***
drvisits    0.250687   0.005329   47.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.875 on 27355 degrees of freedom
  (458946 observations deleted due to missingness)
Multiple R-squared:  0.07484,   Adjusted R-squared:  0.07481 
F-statistic:  2213 on 1 and 27355 DF,  p-value: < 2.2e-16
```

Q2: SL Regression Slope & Intercept Interpretation

- **Slope**
  - A 1 visit increase in the number of doctor visits (drvisits) causes the number of poor mental and physical health days (poorhlth) to increase by an average of 0.251 days
- **Intercept**
  - When the number of doctor visits (drvisits) is zero, the predicted number of poor mental and physical health days (poorhlth) is 3.392 days

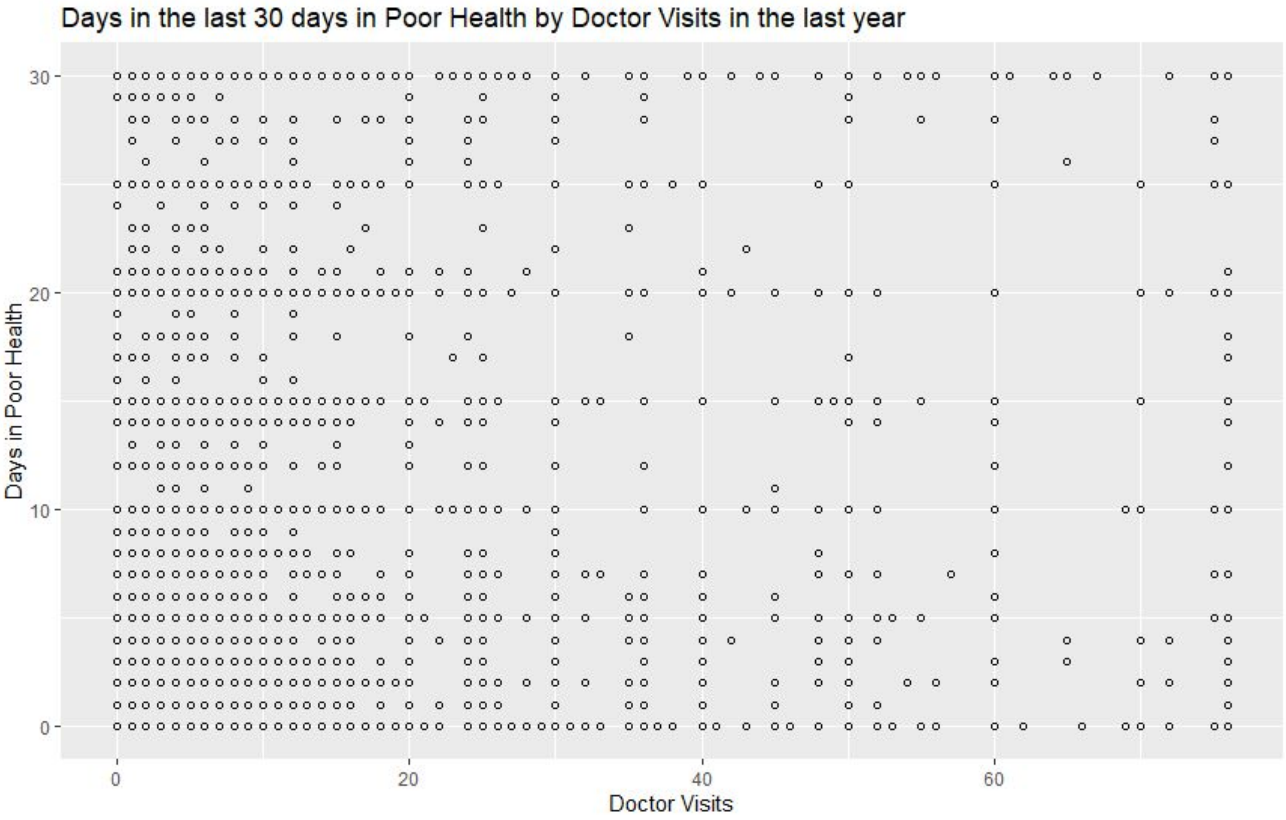
Q2: Simple Linear Regression Significance

- **p-value**
  - The p-value is 2.2×10^{-16}, extremely low.
- **Significance**
  - The regression model is significant because the p-value (2.2×10^{-16}) is lower than the alpha value of 0.05

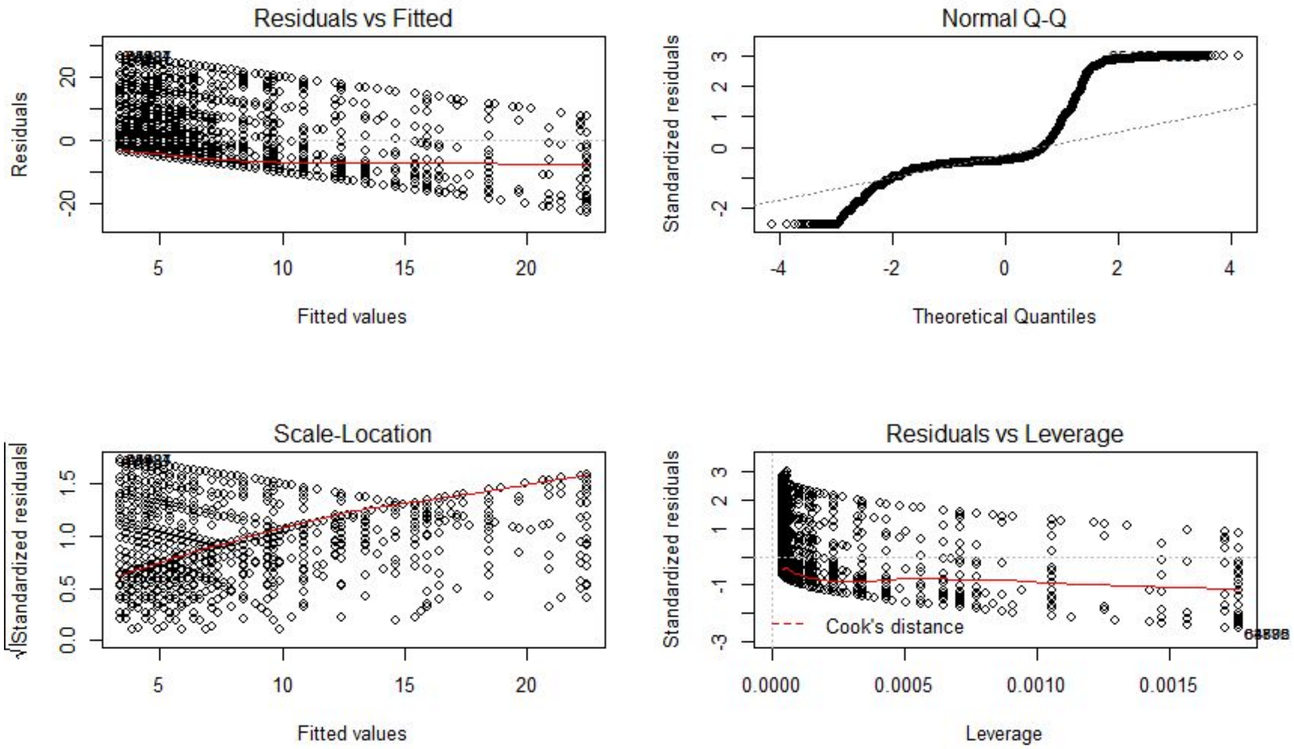
Q2: SL Regression R-Squared

- **R-squared**
  - The adjusted R-squared is 0.075.
- **Interpretation**
  - The fairly low adjusted r-squared value of 0.075 means that the model alone does that accurately predict the poorhlth variable as only about 7.5% of outcomes can be explained by the model.

Q2: SL Regression Visualizations



Q2: SL Regression Diagnostic Charts



Q2: SL Regression Assumption Violations

- **Linearity**
  - Linearity is not violated
- **Homoscedasticity**
  - The homoscedasticity assumption is violated. This is because residuals are constantly decreasing. As a result the subset of data the model is based on may not be correct.
- **Independence**
  - Independence is not violated
- **Normality**
  - The Normality assumption is violated as the Q-Q plot has significant tails on both ends. As a result the model coefficients are not normally distributed and there may be a number of outliers. The model may be skewed as a result.

Research Q3: How well do factors such as race, educational attainment, and the number of days a respondent's mental health was poor predict what age a respondent is diagnosed with cancer?

Variables

There are two categorical independent variables and one continuous variable.

Independent variables:

- *X_RACE:* Race**
- *X_EDUCAG*: Level of educational attainment
- *MENTHLTH:* Number of days the respondent’s mental health was poor

**About 8 different racial options, but the regression was run on three primary races of interest: non-Hispanic white, non-Hispanic Black, and Hispanic.

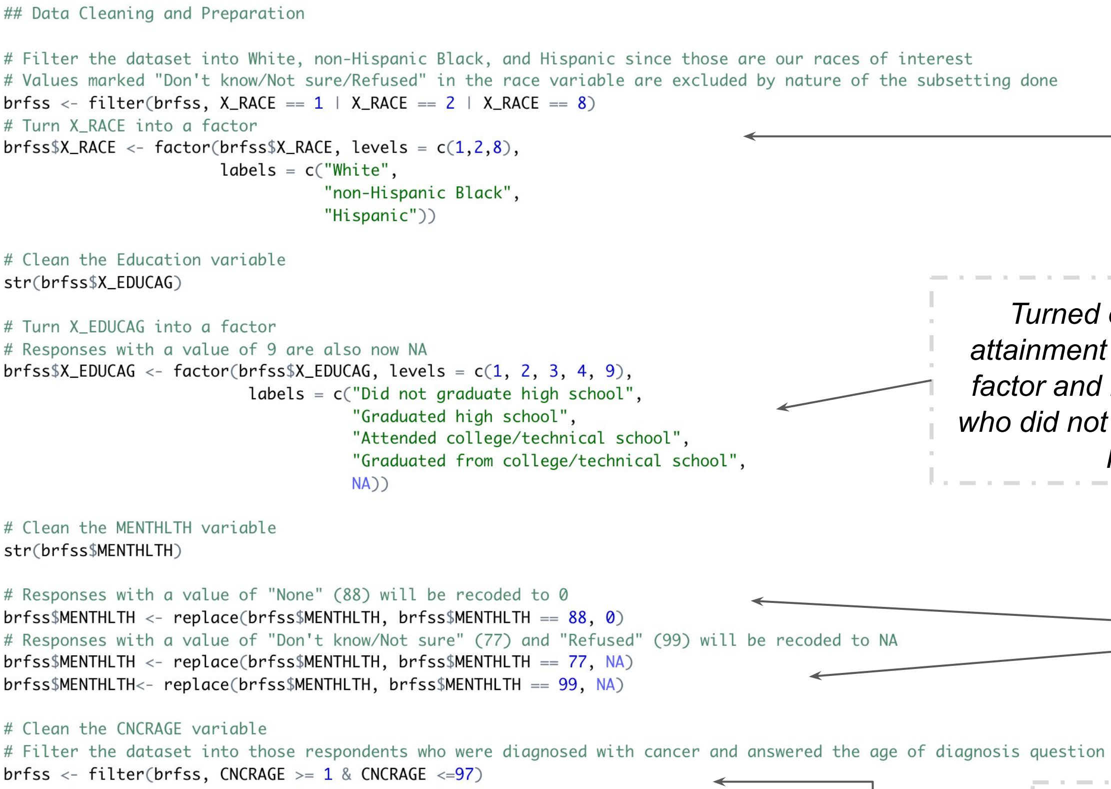
Variables

There is one continuous dependent variable.

Dependent variable:

CNCRAGE: Age of cancer diagnosis in years

Data Preparation and Cleaning



```
## Data Cleaning and Preparation

# Filter the dataset into White, non-Hispanic Black, and Hispanic since those are our races of interest
# Values marked "Don't know/Not sure/Refused" in the race variable are excluded by nature of the subsetting done
brfss <- filter(brfss, X_RACE == 1 | X_RACE == 2 | X_RACE == 8)
# Turn X_RACE into a factor
brfss$X_RACE <- factor(brfss$X_RACE, levels = c(1,2,8),
                       labels = c("White",
                                  "non-Hispanic Black",
                                  "Hispanic"))

# Clean the Education variable
str(brfss$X_EDUCAG)

# Turn X_EDUCAG into a factor
# Responses with a value of 9 are also now NA
brfss$X_EDUCAG <- factor(brfss$X_EDUCAG, levels = c(1, 2, 3, 4, 9),
                         labels = c("Did not graduate high school",
                                    "Graduated high school",
                                    "Attended college/technical school",
                                    "Graduated from college/technical school",
                                    NA))

# Clean the MENTHLTH variable
str(brfss$MENTHLTH)

# Responses with a value of "None" (88) will be recoded to 0
brfss$MENTHLTH <- replace(brfss$MENTHLTH, brfss$MENTHLTH == 88, 0)
# Responses with a value of "Don't know/Not sure" (77) and "Refused" (99) will be recoded to NA
brfss$MENTHLTH <- replace(brfss$MENTHLTH, brfss$MENTHLTH == 77, NA)
brfss$MENTHLTH<- replace(brfss$MENTHLTH, brfss$MENTHLTH == 99, NA)

# Clean the CNCRAGE variable
# Filter the dataset into those respondents who were diagnosed with cancer and answered the age of diagnosis question
brfss <- filter(brfss, CNCRAGE >= 1 & CNCRAGE <=97)
```

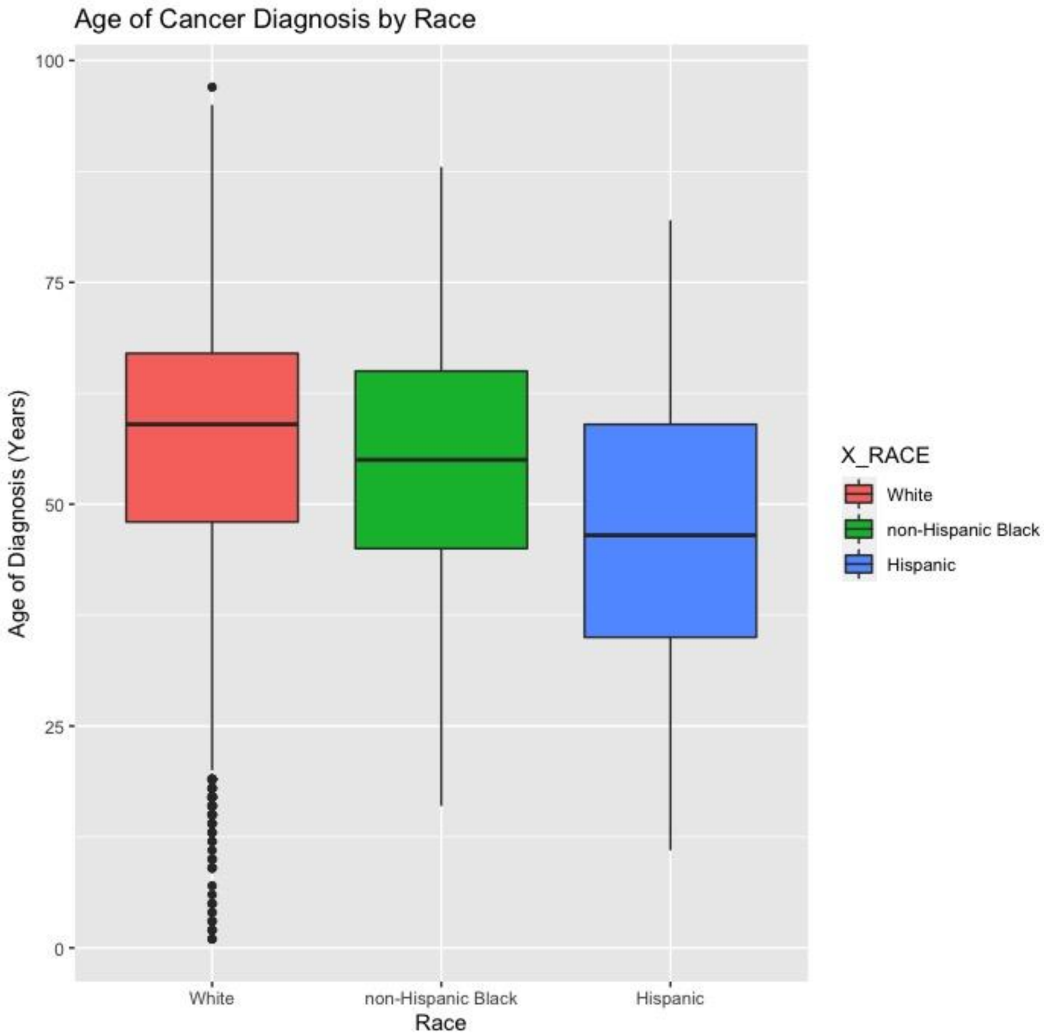
Filtered the dataset into those who responded to our races of interest

Turned educational attainment variable into a factor and recoded those who did not responded into NAs

Those who reported no poor mental health days were recoded from “None” to a value of 0. Those who refused to respond or were not sure were recoded into NAs

Those who refused to answer for the age of their cancer diagnosis were recoded to NAs.

Q3: Visualizations

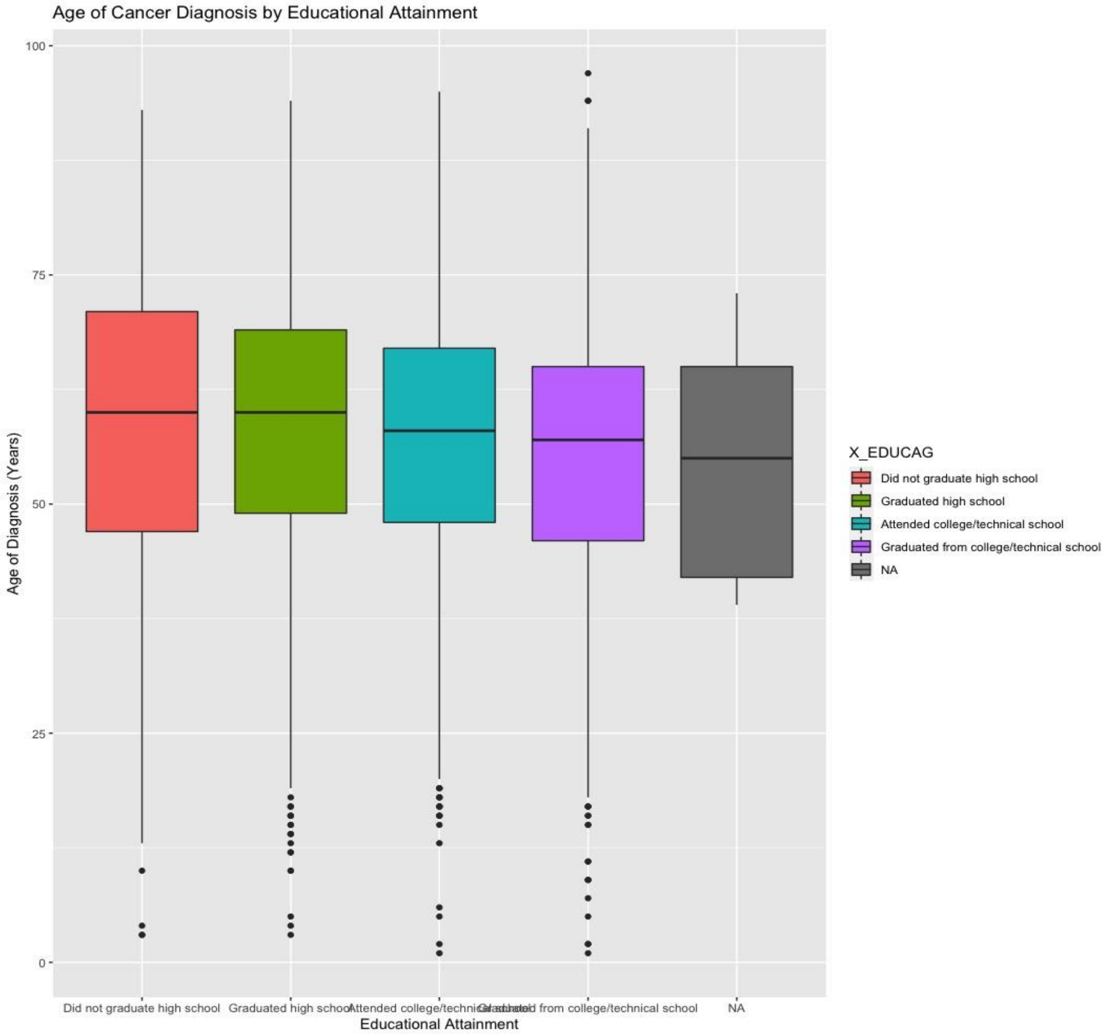


See that the average age of diagnosis for white respondents is higher than that of the other races.

Supported by research by national health institutes who have indicated that minorities are more at risk for cancer.

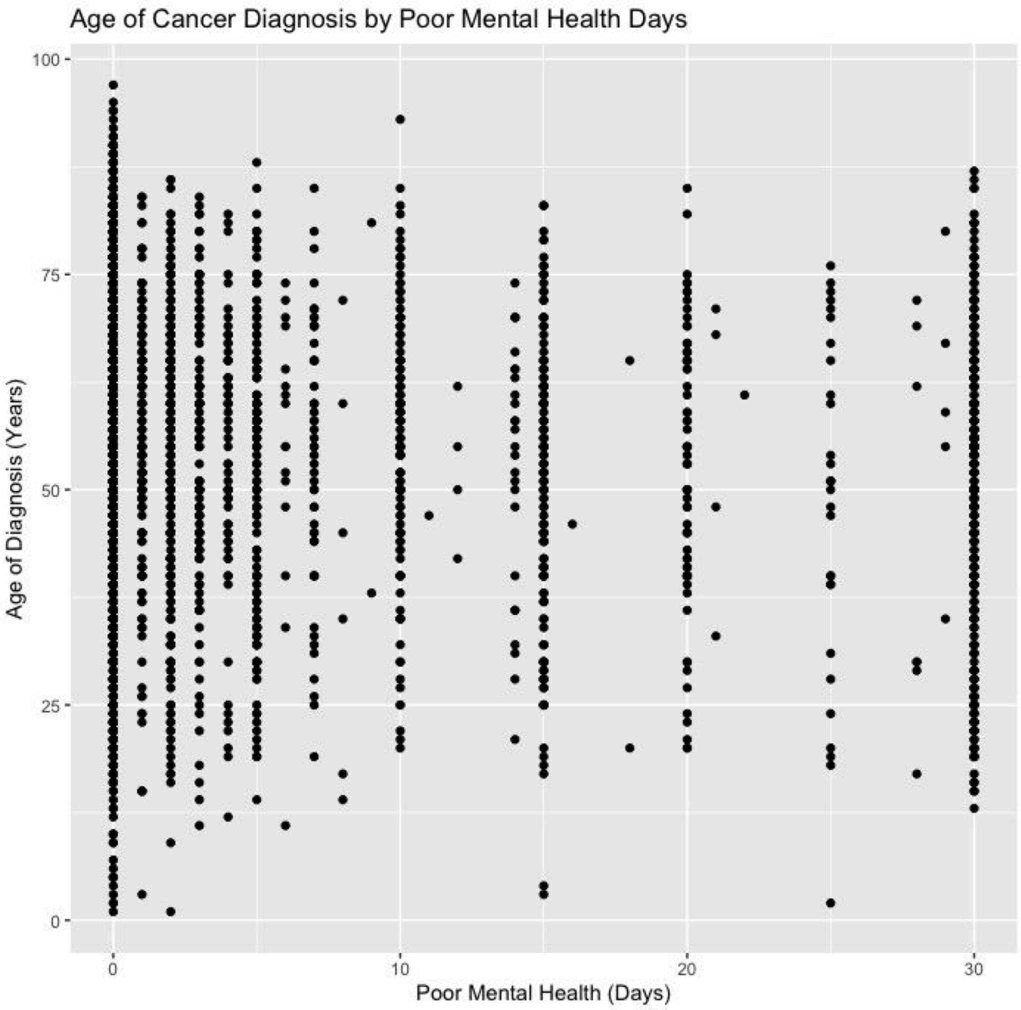
See that the average age of diagnosis for all the educational levels is generally the same, hovering around 60 years old.

The middle 50% for those who refused to answer the question does dip lower than the rest though.

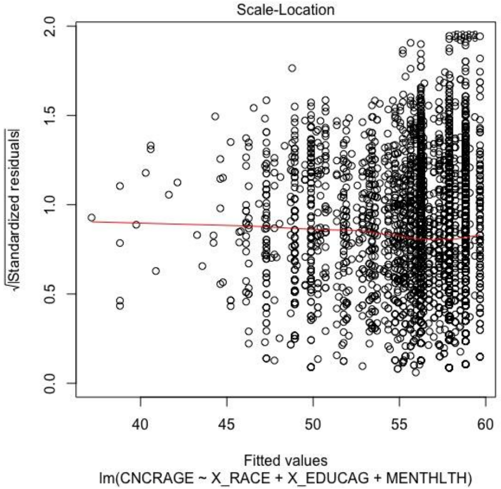


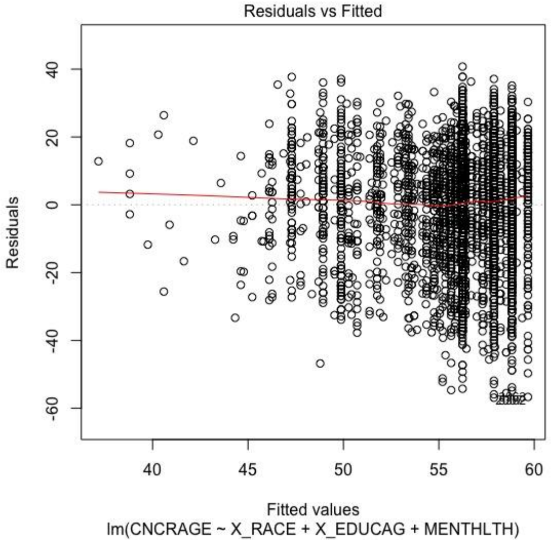
Don’t see any true trend here. However, notice that the greatest range of ages come from those who reported no poor mental health days and those who reported the max of 30 days.

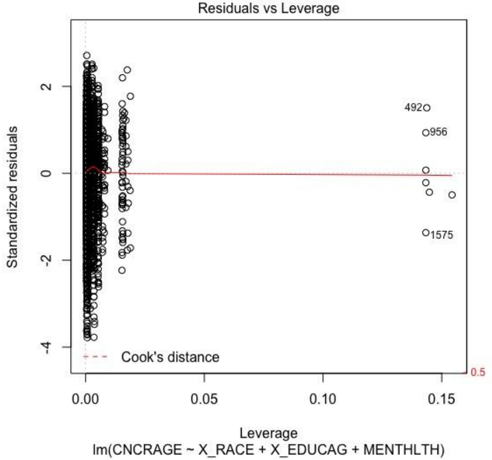
More people reported having a number of mental health days less than half of the month.

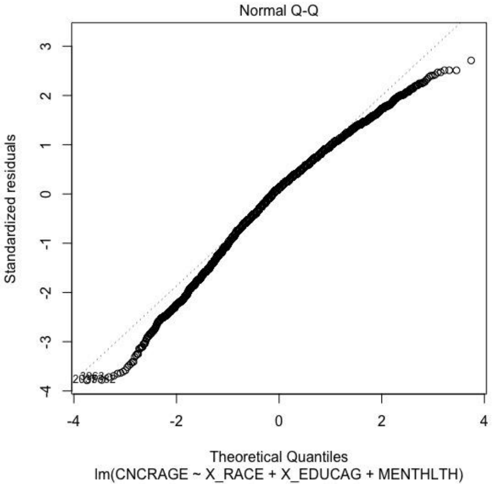


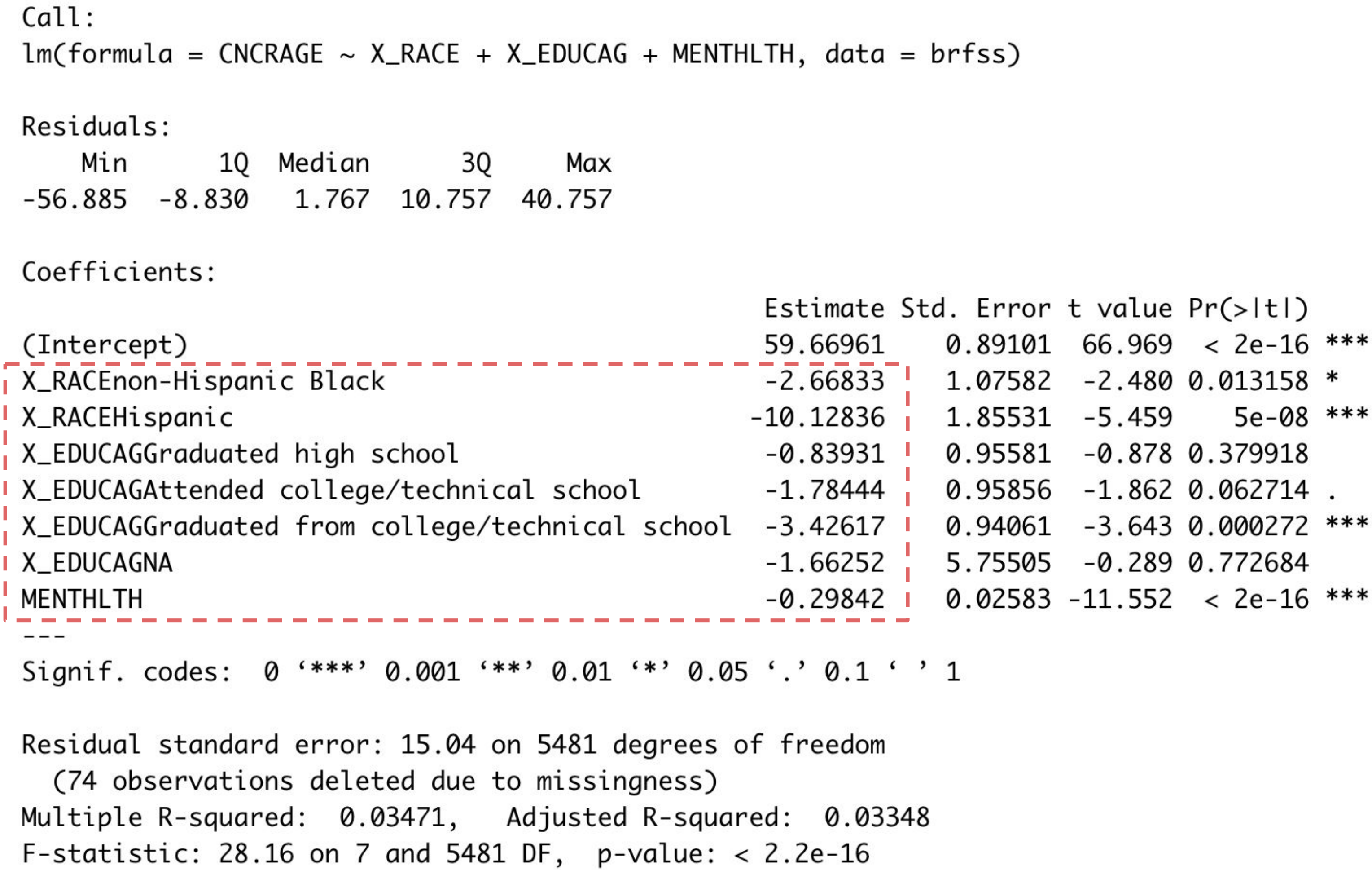
Q3: Assumptions











```
Call:
lm(formula = CNCRAGE ~ X_RACE + X_EDUCAG + MENTHLTH, data = brfss)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.885  -8.830   1.767  10.757  40.757 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              59.66961    0.89101  66.969  < 2e-16 ***
X_RACEnon-Hispanic Black                 -2.66833    1.07582  -2.480 0.013158 *  
X_RACEHispanic                          -10.12836    1.85531  -5.459    5e-08 ***
X_EDUCAGGraduated high school            -0.83931    0.95581  -0.878 0.379918    
X_EDUCAGAttended college/technical school -1.78444    0.95856  -1.862 0.062714 .  
X_EDUCAGGraduated from college/technical school -3.42617    0.94061  -3.643 0.000272 ***
X_EDUCAGNA                               -1.66252    5.75505  -0.289 0.772684    
MENTHLTH                                 -0.29842    0.02583 -11.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.04 on 5481 degrees of freedom
  (74 observations deleted due to missingness)
Multiple R-squared:  0.03471,	Adjusted R-squared:  0.03348 
F-statistic: 28.16 on 7 and 5481 DF,  p-value: < 2.2e-16
```

Q3: Interpreting Slope Coefficients

By race

non-Hispanic Black

Hispanic

- The average age of cancer diagnosis is 2.668 years lower in those who identify as *non-Hispanic Black* when compared to those who identify as non-Hispanic white, holding all other variables constant.
- The average age of cancer diagnosis is 10.128 years lower in those who identify as *Hispanic* when compared to those who identify as non-Hispanic white, holding all other variables constant.

Q3: Interpreting Slope Coefficients

By Educational Attainment

Graduated High School

Attended some college

Graduated from College

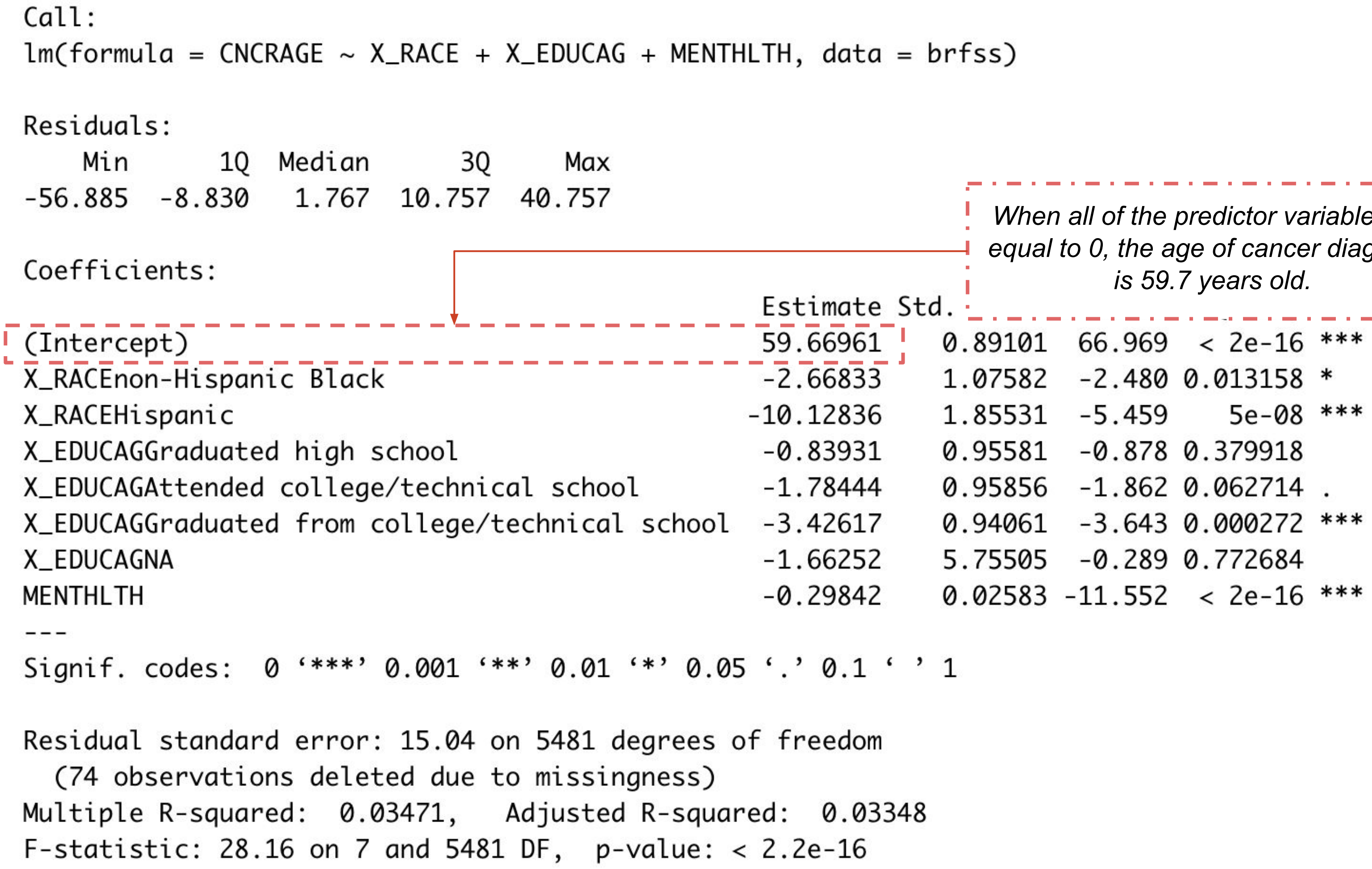
Refused to answer

- The average age of cancer diagnosis is 0.839 years lower for those who *graduated high school* compared to those who did not, holding all other variables constant.
- The average age of cancer diagnosis is 1.784 years lower for those who *attended some college* when compared to those who did not graduate from high school, holding all other variables constant.
- The average age of cancer diagnosis is 3.426 years lower for those who *graduated from college* when compared to those who did not graduate from high school, holding all other variables constant.
- The average age of cancer diagnosis is 1.66 years lower for those who *refused to answer the question* about educational attainment when compared to those who did not graduate from high school, holding all other variables constant.

Q3: Interpreting Slope Coefficients

By Poor Mental Health Days

- For every additional day a respondent’s mental health was poor, the age of cancer diagnosis decreases 0.298 years, holding all other variables constant.



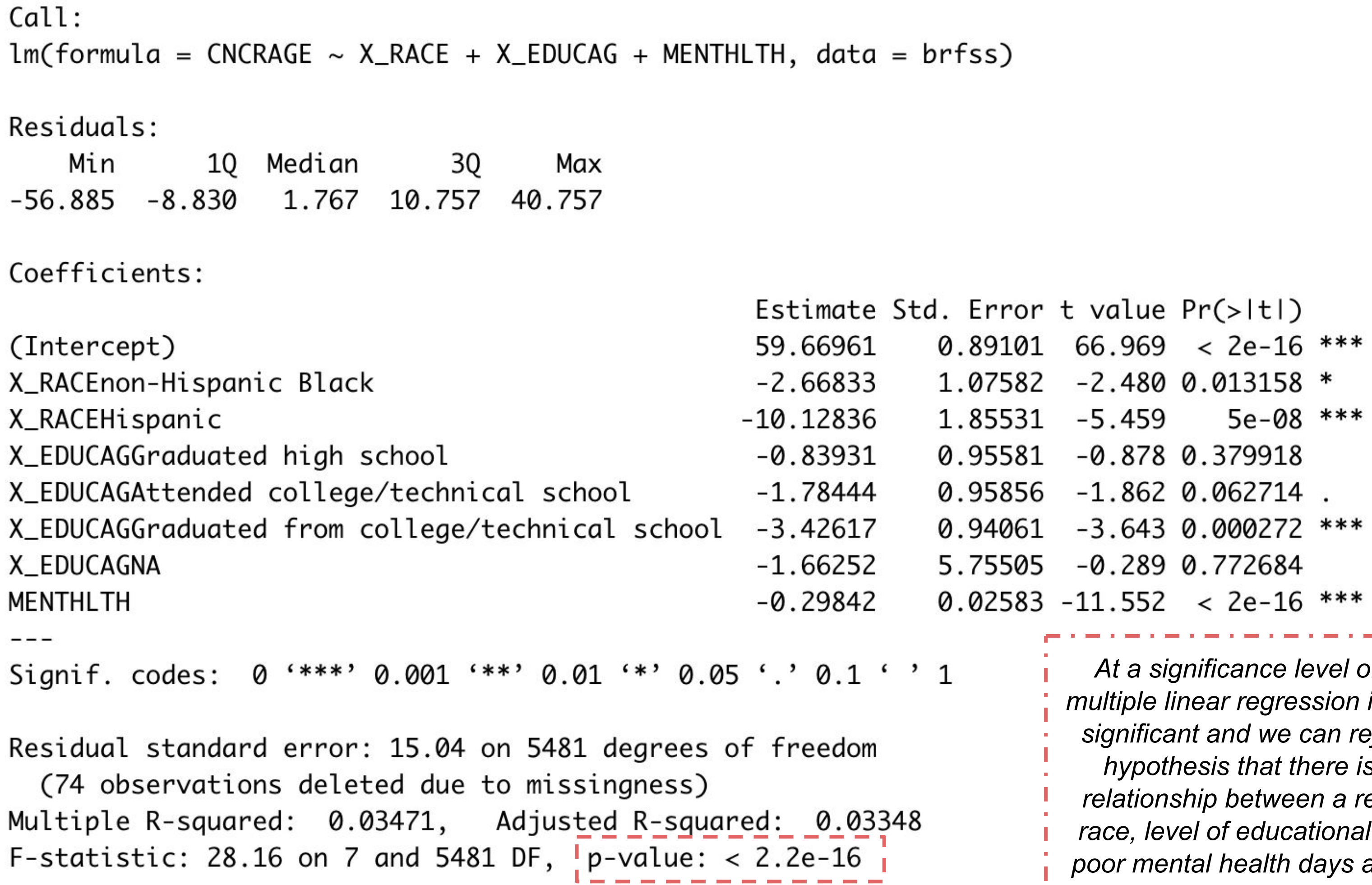
```
Call:
lm(formula = CNCRAGE ~ X_RACE + X_EDUCAG + MENTHLTH, data = brfss)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.885  -8.830   1.767  10.757  40.757 

Coefficients:
                                             Estimate Std.
(Intercept)                                  59.66961    0.89101  66.969  < 2e-16 ***
X_RACEnon-Hispanic Black                     -2.66833    1.07582  -2.480 0.013158 *  
X_RACEHispanic                              -10.12836    1.85531  -5.459    5e-08 ***
X_EDUCAGGraduated high school                -0.83931    0.95581  -0.878 0.379918    
X_EDUCAGAttended college/technical school    -1.78444    0.95856  -1.862 0.062714 .  
X_EDUCAGGraduated from college/technical school -3.42617    0.94061  -3.643 0.000272 ***
X_EDUCAGNA                                   -1.66252    5.75505  -0.289 0.772684    
MENTHLTH                                     -0.29842    0.02583 -11.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.04 on 5481 degrees of freedom
  (74 observations deleted due to missingness)
Multiple R-squared:  0.03471,	Adjusted R-squared:  0.03348 
F-statistic: 28.16 on 7 and 5481 DF,  p-value: < 2.2e-16
```

When all of the predictor variables are equal to 0, the age of cancer diagnosis is 59.7 years old.



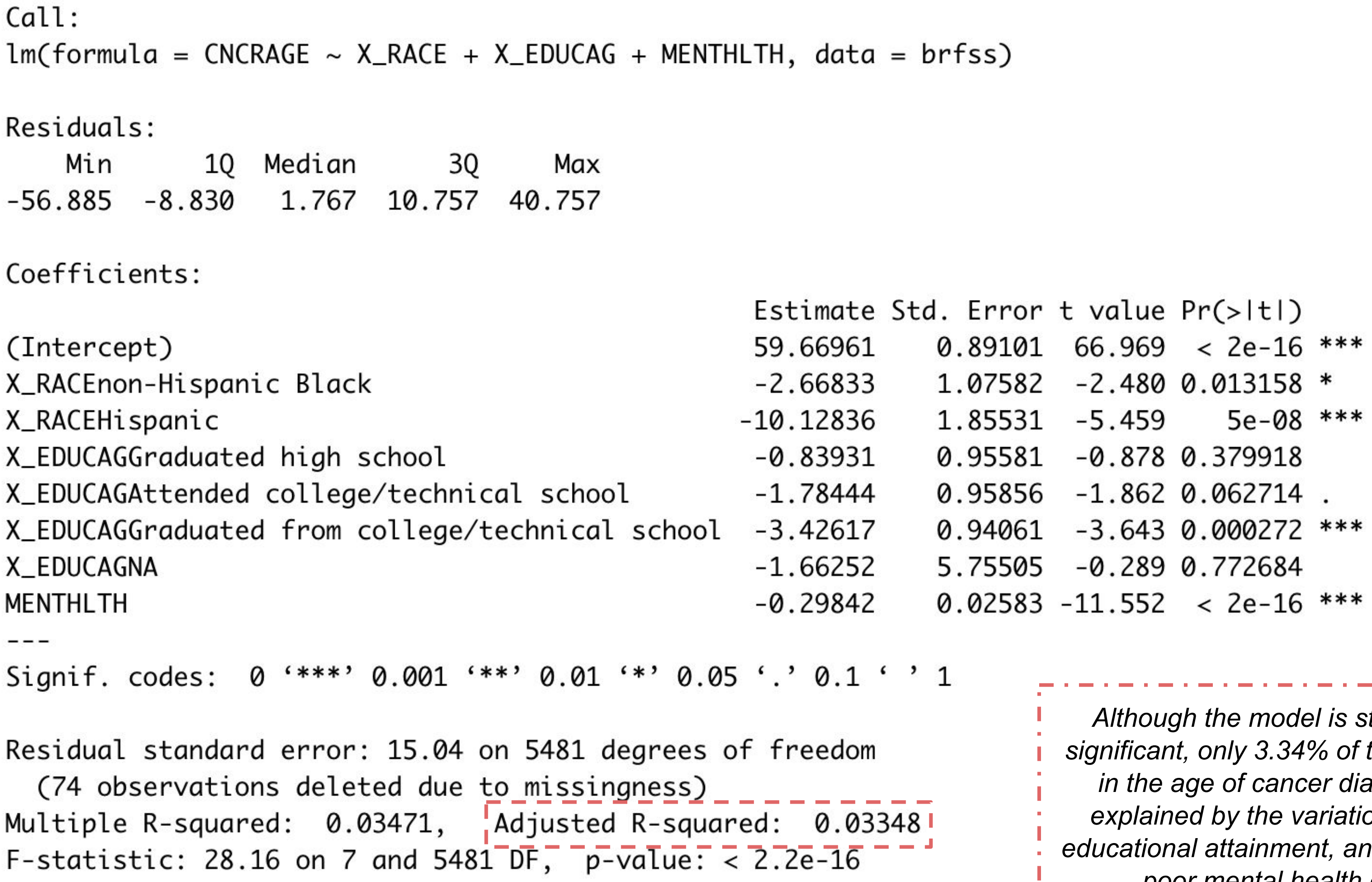
```
Call:
lm(formula = CNCRAGE ~ X_RACE + X_EDUCAG + MENTHLTH, data = brfss)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.885  -8.830   1.767  10.757  40.757 

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                      59.66961    0.89101  66.969  < 2e-16 ***
X_RACEnon-Hispanic Black                         -2.66833    1.07582  -2.480 0.013158 *  
X_RACEHispanic                                  -10.12836    1.85531  -5.459    5e-08 ***
X_EDUCAGGraduated high school                    -0.83931    0.95581  -0.878 0.379918    
X_EDUCAGAttended college/technical school        -1.78444    0.95856  -1.862 0.062714 .  
X_EDUCAGGraduated from college/technical school  -3.42617    0.94061  -3.643 0.000272 ***
X_EDUCAGNA                                       -1.66252    5.75505  -0.289 0.772684    
MENTHLTH                                         -0.29842    0.02583 -11.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.04 on 5481 degrees of freedom
  (74 observations deleted due to missingness)
Multiple R-squared:  0.03471,	Adjusted R-squared:  0.03348 
F-statistic: 28.16 on 7 and 5481 DF,  p-value: < 2.2e-16
```

At a significance level of 0.05, this multiple linear regression is statistically significant and we can reject the null hypothesis that there is no linear relationship between a respondent’s race, level of educational attainment, poor mental health days and their age of cancer diagnosis.



```
Call:
lm(formula = CNCRAGE ~ X_RACE + X_EDUCAG + MENTHLTH, data = brfss)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.885  -8.830   1.767  10.757  40.757 

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                      59.66961    0.89101  66.969  < 2e-16 ***
X_RACEnon-Hispanic Black                         -2.66833    1.07582  -2.480 0.013158 *  
X_RACEHispanic                                  -10.12836    1.85531  -5.459    5e-08 ***
X_EDUCAGGraduated high school                    -0.83931    0.95581  -0.878 0.379918    
X_EDUCAGAttended college/technical school        -1.78444    0.95856  -1.862 0.062714 .  
X_EDUCAGGraduated from college/technical school  -3.42617    0.94061  -3.643 0.000272 ***
X_EDUCAGNA                                       -1.66252    5.75505  -0.289 0.772684    
MENTHLTH                                         -0.29842    0.02583 -11.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.04 on 5481 degrees of freedom
  (74 observations deleted due to missingness)
Multiple R-squared:  0.03471,	Adjusted R-squared:  0.03348 
F-statistic: 28.16 on 7 and 5481 DF,  p-value: < 2.2e-16
```

Although the model is statistically significant, only 3.34% of the variation in the age of cancer diagnosis is explained by the variation in race, educational attainment, and number of poor mental health days.

Research Q4: Is the average BMI of adults with diabetes higher than the average BMI of adults without diabetes?

- **Associated variables:**
    - DIABETE3
        - Categorical variable with “has diabetes” and “does not have diabetes”
    - _BMI5
        - Continuous variable with the computed Body Mass Index

Q4: Data Cleaning

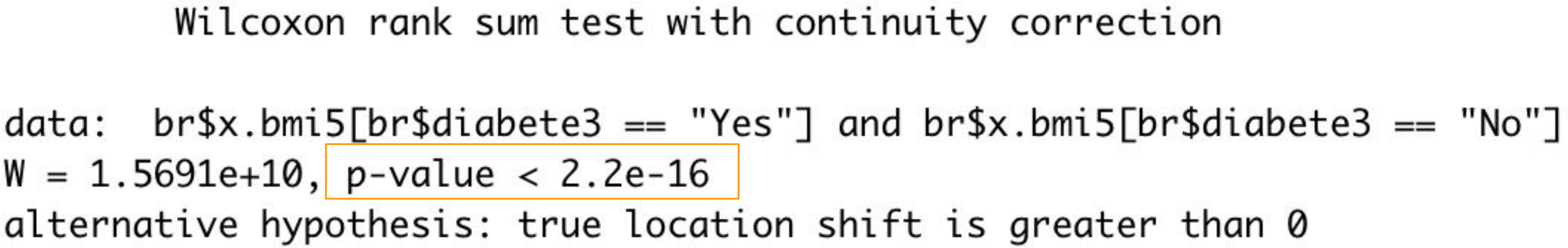
- Filter DIABETE3 values with 1 and 3, representing “has diabetes” and “does not have diabetes”
- Recode 1 to “Yes” and 3 to “No”
- Filter for BMI with responses and within valid range

Null Hypothesis = The mean BMI for those with diabetes is less than for those without diabetes.

Alternative Hypothesis = The mean BMI for those with diabetes is greater than those without diabetes.

Alpha = 0.05

Type of test (upper, lower, two-sided) = Upper-Tail

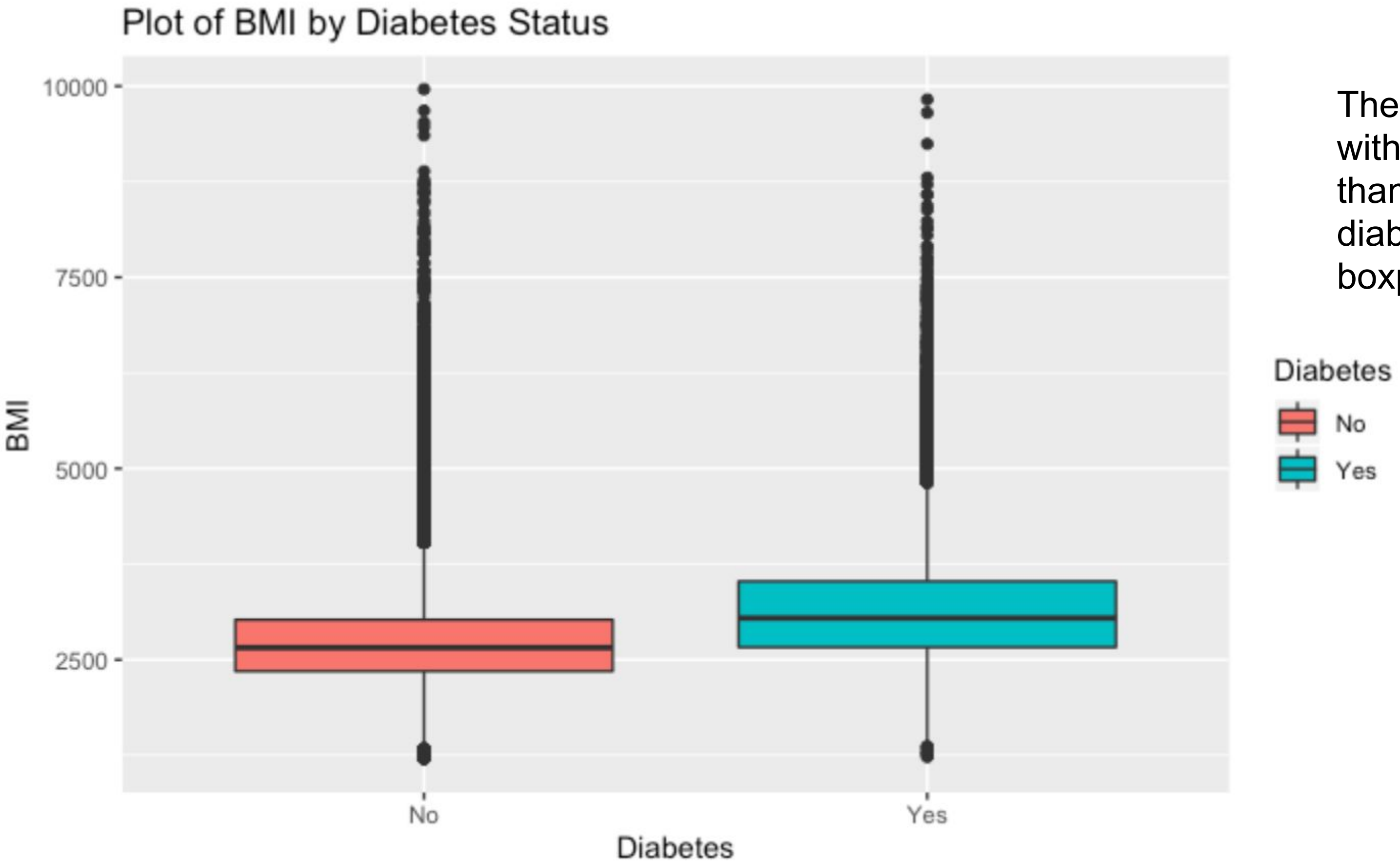


```
        Wilcoxon rank sum test with continuity correction

data:  br$x.bmi5[br$diabete3 == "Yes"] and br$x.bmi5[br$diabete3 == "No"]
W = 1.5691e+10, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

We reject the null hypothesis because our **p-value** (<2.2e-16) is less than our **alpha** of 0.05.

Conclusion = At a significance level of 0.05, we reject the null hypothesis and conclude that there is statistically significant evidence that the mean BMI of those with diabetes is greater than those without diabetes.



The mean BMI for those with diabetes is higher than those without diabetes as shown by the boxplot.