Final Project

The Unemployed Guide to Living Forever: A Lifestyle Data Analysis

Spring 2024 PSTAT 126: Regression Analysis

Richard Zhang, Ryan Sohn, Oscar Baek, Corbin White, Jeffrey Chen

Dataset and Group Tasks

Role Distribution

Name	Responsibilities
Richard Zhang	Tasked with creating preliminary insights regarding our data and tasked with editing and reviewing the coding segments of each team member's section. Also responsible for completing the written portion of the report, coordinating team workflows, and scheduling meetings in an Agile system.
Ryan Sohn	Tasked with interpreting our linear models and conducting hypothesis testing on the correlation and slope of our simple linear regression model.
Oscar Baek	Tasked with performing residual analysis to help confirm model assumptions and providing insights related to interpretation of residuals.
Corbin White	Responsible for data cleaning and generating early insights from the dataset to guide our modeling approach.
Jeffrey Chen	Tasked with building visualizations to support data interpretation. Adds annotations and comments to improve clarity and accessibility of plots.

Data Information

Data Name/Title: Human Age Prediction Synthetic Dataset

Author/Owner: M Abdullah and Shahzaib Yaqoob

Date of Publication: *August 2024*

Publication Venue: Kaggle

Retrieval Date: October 24, 2024

Link: https://www.kaggle.com/datasets/abdullah0a/human-age-prediction-synthetic-dataset

Initial Insights

With the cultural rise and emphasis on mental and physical health, we as active unemployed human beings want to gain better insights on how we can improve/maintain our quality of life. In analyzing this data, we intend to leverage the variables to understand how certain variables such as gender or education will impact one's emotional and mortal being. We are also interested in finding out if people with great income and education live happy and healthy lives. As many of us have been pressured to try our best to get an amazing education and work in fields that are high-paying to live "great" lives from family and peers. With this data, we can resolve this myth. So variables we will need to look at to test this myth would be stress levels, cognitive function, education, diet, smoking status, alcohol consumption, income level, family history, chronic diseases, physical activity level, and mental health status. With this dataset, we plan to understand the following categorical data: gender and education as well. We will also be utilizing the ensuing quantitative variables such as bone density, hearing ability, and cognitive function. By quick observation, we can make multiple assumptions such as blood pressure(s/d), cholesterol levels(mg/dL), and BMI having influences on weight(kg). We can also make an assumption that age(yr) impacts pollution exposure, education level, physical activity level, and bone density(g/emˆ2).

Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

Research Questions

Question 1:

In society, we often hear about the brain's dependence on a healthy body, but is this true? Is there a link between cognitive performance and physical health metrics like vision clarity, bone density, blood glucose levels, and the natural aging process? Are individuals with better physical health markers more likely to maintain sharp cognitive abilities, or do other factors come into play? A prime example for this dilemma is Stephen Hawking. This study aims to explore these relationships, shedding light on how our physical and mental health interact as we age.

Question 2

What is the impact of lifestyle choices (physical activity, smoking, alcohol consumption, etc) on age prediction? Have you ever realized that it was always rude to ask a lady their age? So most men and women tend to try to guess it. And most of the times we are inaccurate. This is also because we base it off of one's appearance. Which make it difficult to accurately predict one's age. Well, we "The Unemployed" are on the mission to find out what lifestyle choices impact one's predictions on another age. As one's health and actions towards their body will affect their appearance and affects one's ability to predict their age.

Hypotheses

Hypothesis 1: Physical Health metrics such as vision sharpness, bone density, blood glucose levels, and age do not have any significant relationship with cognitive function.

Hypothesis 2: Lifestyle choices such as alcohol consumption, smoking, and physical activity do not have any correlation with one's age.

Exploratory Data Analysis (EDA)

Data Cleaning

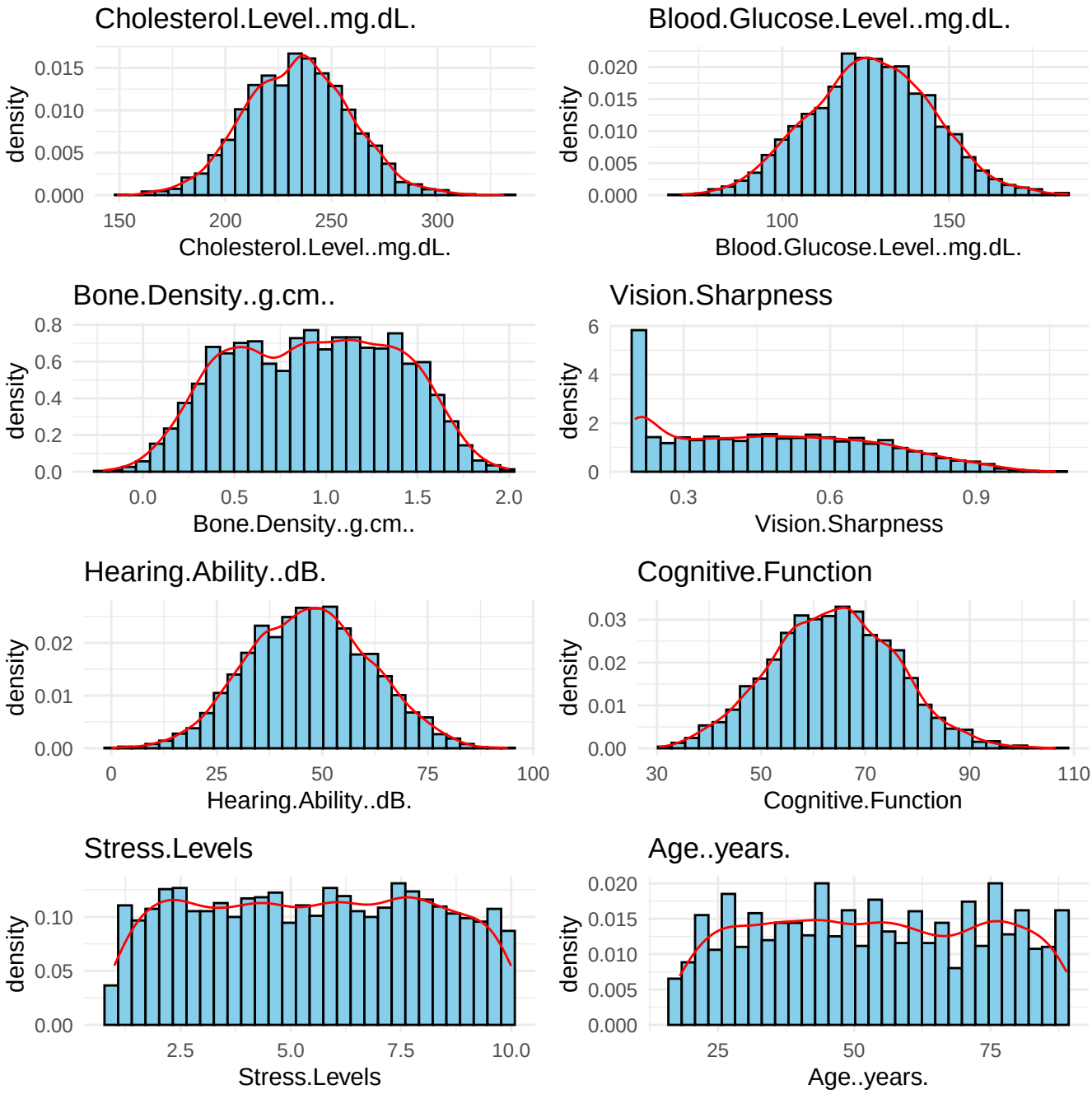
Checking for any missing values in out dataset:

```
##                      Gender        Blood.Pressure..s.d.
##                           0                           0
##   Cholesterol.Level..mg.dL. Blood.Glucose.Level..mg.dL.
##                           0                           0
##        Bone.Density..g.cm..            Vision.Sharpness
##                           0                           0
##        Hearing.Ability..dB.     Physical.Activity.Level
##                           0                           0
##              Smoking.Status         Alcohol.Consumption
##                           0                           0
##                        Diet          Cognitive.Function
##                           0                           0
##               Stress.Levels                 Age..years.
##                           0                           0
```

- Since there are no missing values in out dataset, it is not necessary to remove any data entries.

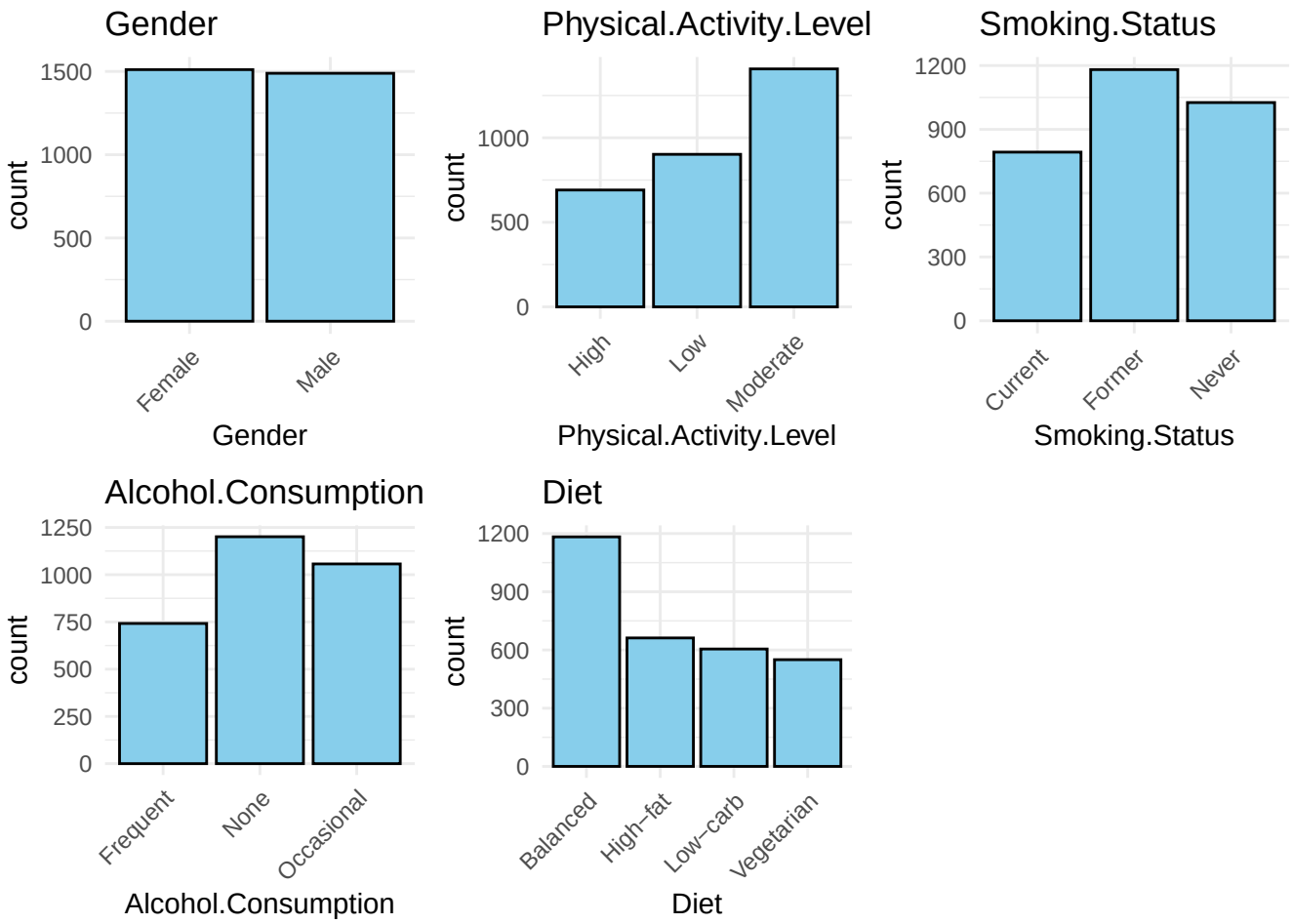
Data Visualization

Histogram matrix of relevant numerical variables to help give insight on out data:



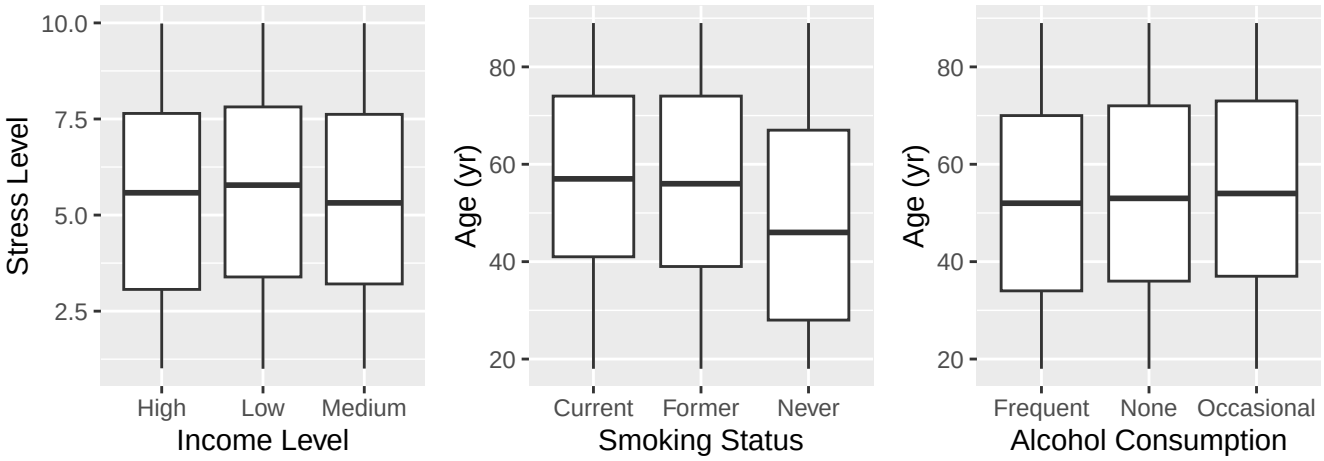
- We can see that factors such as Cholesterol Levels, Blood Glucose Levels, Hearing Ability, and Cognitive Function appear to be normally distributed.
- The Vision Sharpness bargraph is interesting to look at. As we can see that it skewed to the right. However, it goes flat in the middle and then slips down.
- We can also observe that Stress Levels are the most evenly distributed factor that is relevant to us.

Histogram matrix of relevant categorical variables:



- Based on the bargraphs of our categorical variables, we can see that most of graphs are what we can expect. However, one that is interesting to look at is the fact that out data has lots of former smokers. And that many have smoked or are currently smoking right now (same with alcohol consumption).

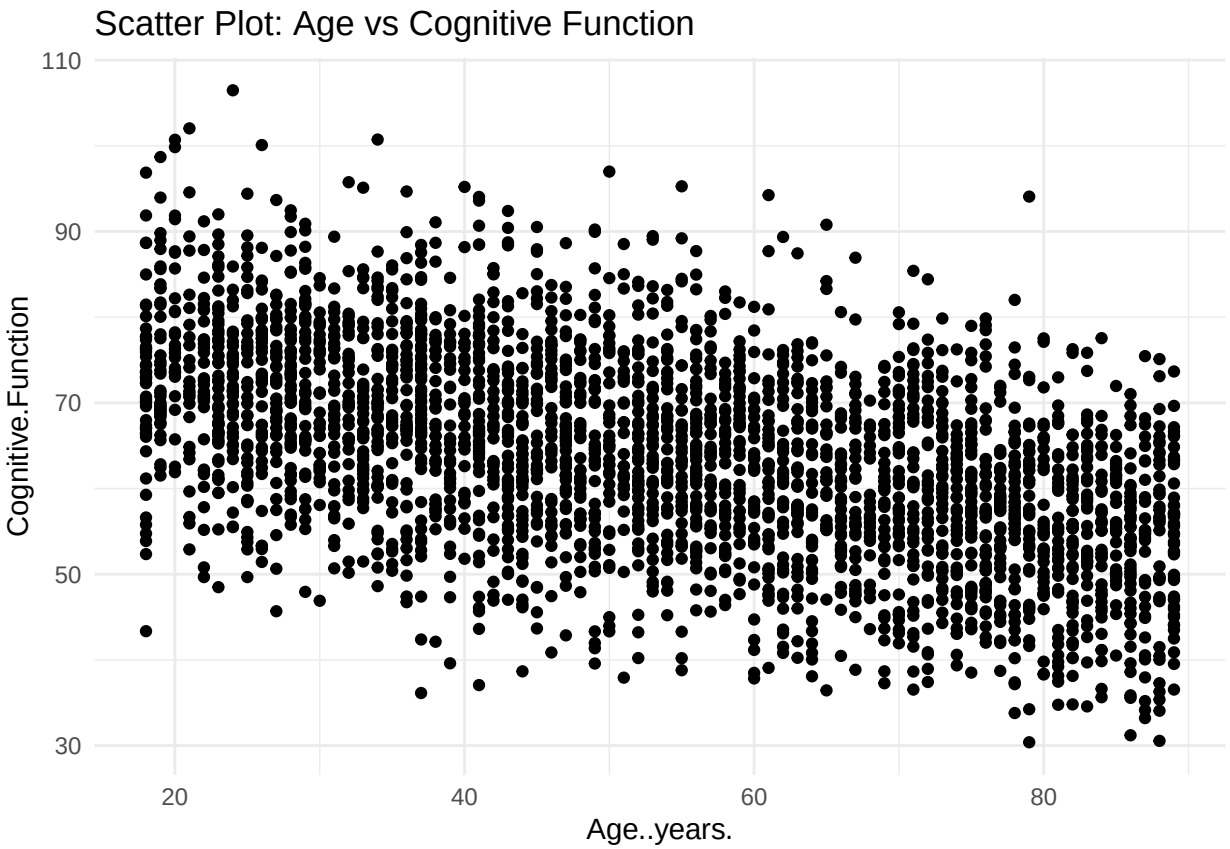
Boxplot of Relevant Variables:



- Based on this boxplot, we can see that we do not have any outliers. This can be because responses for stress levels are only range from 0 to 10. However, it is interesting to see that the median stress levels for all income levels are pretty similar. This is very relevant as this is implying that no matter how much you make it does not affects one's stress levels.
- For the smoking and age boxplot, we can see that there are not outliers. We can see that the results are not that surprising due to the fact that people who never smoked are younger. As the older the person gets, the likelihood of them smoking is greater with more life experiences.

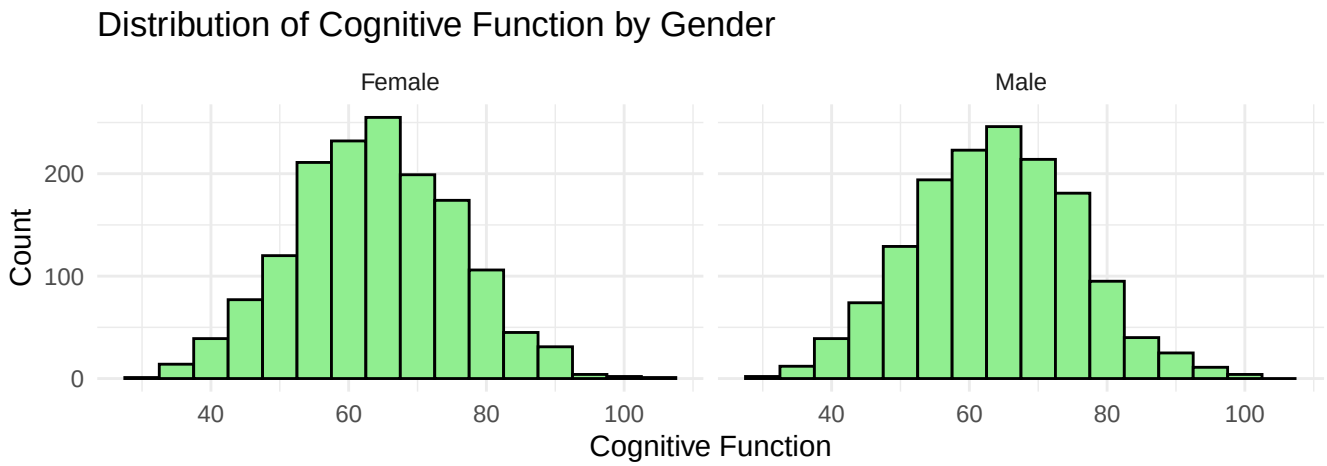
*It is interesting to see that a the median age of people who have not consumed alcohol are in their 50s with is extremely surprising.

Scatter plot for Age vs Cognitive Function:



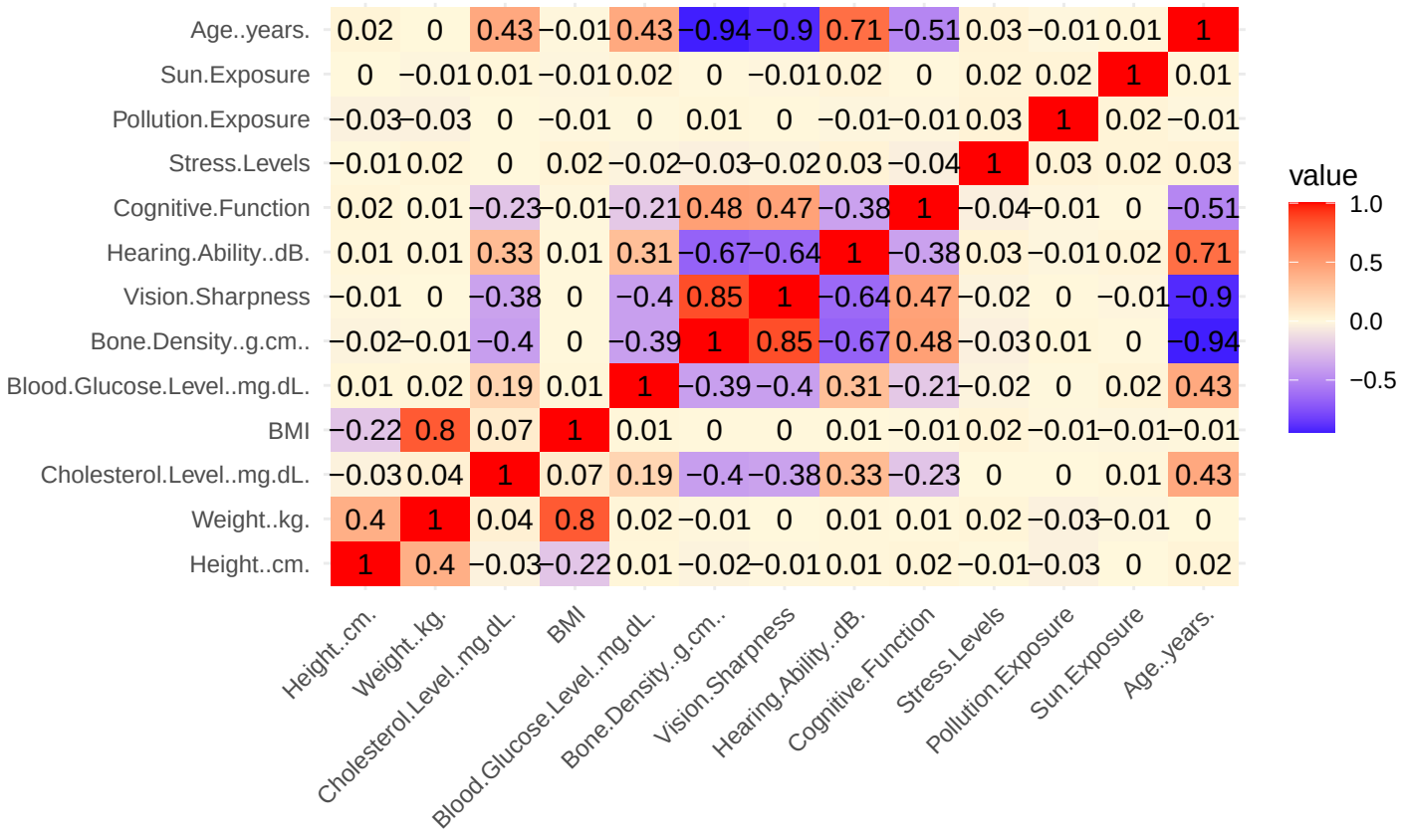
- Based on the following scatterplot, we can see a clear negative linear relation between Cognitive Function and Age. Though many of the points are cluster together, if we were to draw an imaginary line following the points we can see a negative linear relation between the two variables. We can confirm this statement later when we look at their correlations.

Facted Histogram of Cognitive Function by Gender:



- Based of the faceted histogram with gender and cognitive function, we can see that both genders are normally distributed. With pretty identical distributions.

Correlation Heatmap of our Dataset:



- It is very interesting to see that there are so many relations between the dataset variables that have any relevant correlation between them. We can see that we have significant associations between our variables for our first research question.
- Like in the scatterplot, we can see that there is a significant negative correlation between Cognitive

Function and Age.

Skewness and Kurtosis:

```
## Skewness of Age: 0.02465048
## Kurtosis of Age: -1.187818
## Skewness of Cognitive Function: 0.05391167
## Kurtosis of Cognitive Function: -0.1696383
## Skewness of Stress Level: 0.003200952
## Kurtosis of Stress Level: -1.190207
```

- Based on the following results, we can see that the distribution of all three variables above are skewed to the right. With Cognitive Function having the larger skewness.
- We can see that the kurtosis of all three variables have flat/spread out distribution. Which means that there are little to no outliers as well as most of our data entries are evenly distributed across the range.

Initial Insight

Based off all the visualizations done above, it was extremely interesting to see that most of our data entries have or had a history of smoking and consuming alcohol. This is going to be extremely interesting to find out if the two factors will affect age. As we know smoking and consuming alcohol can make a person's appearance look older. A notable find is that most of our relevant variables that we are interested in studying are normally distributed. Another thing that is fascinating is that we do not have any extreme values that will influence our conclusions when we conduct hypothesis testing.

Regression Analysis and Interpretation:

Cognitive Health Research (Research Question 1)

We will first be conducting our Regression Analysis on our first research topic regarding Cognitive Function and physical health metrics.

Simple and Mulitple Linear Regression

Simple Linear Model

```
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years., data = health_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.523  -6.828  -0.048   6.948  37.615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.39842    0.51517   154.1   <2e-16 ***
## Age..years. -0.29036    0.00899   -32.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 2998 degrees of freedom
## Multiple R-squared:  0.2581, Adjusted R-squared:  0.2579
## F-statistic:  1043 on 1 and 2998 DF,  p-value: < 2.2e-16
```

- The regression equation for this linear model can be expressed as:

$Cognitive\ Function = 70.30842 - 0.29036 * Age + \epsilon$

- We can see that the baseline Cognitive Function that people who responded to the survey reported when our Age predictor is zero is 70.30842 from a score range of 0 to 100 (Keep in mind that no such state exists).
- Based on the model, we can see that as age increases by one year cognitive function decreases by 0.29036. This effect is highly statistically significant (p-value $<2 \times 10^{-16}$).

-We can also see that our overall model is extremely significant (p-value $<2 \times 10^{-16}$)

- We can also see that our R^2 is 0.2581. Which implies that approximately 25.81 of the variability in Cognitive Function can be explained by Age. However, after adjusting the number of ages, the model explains about 25.79 of the variability in Cognitive Function. Since our R^2 value is 0.2581, it also suggests that our model is a poor fit.

Multiple Regression Model

```
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years. + Hearing.Ability..dB. +
##     Vision.Sharpness + Bone.Density..g.cm.. + Blood.Glucose.Level..mg.dL. +
##     Cholesterol.Level..mg.dL., data = health_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.684  -6.832  -0.078   6.921  36.795
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 78.227262   3.737917  20.928  < 2e-16 ***
## Age..years.                 -0.249081   0.033518  -7.431  1.4e-13 ***
## Hearing.Ability..dB.        -0.038137   0.018384  -2.074   0.0381 *
## Vision.Sharpness             2.731449   2.020719   1.352   0.1766
## Bone.Density..g.cm..        -0.140770   1.200434  -0.117   0.9067
## Blood.Glucose.Level..mg.dL.  0.007609   0.011236   0.677   0.4983
## Cholesterol.Level..mg.dL.   -0.005864   0.008367  -0.701   0.4834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.12 on 2993 degrees of freedom
## Multiple R-squared:  0.2599, Adjusted R-squared:  0.2584
## F-statistic: 175.1 on 6 and 2993 DF,  p-value: < 2.2e-16
```

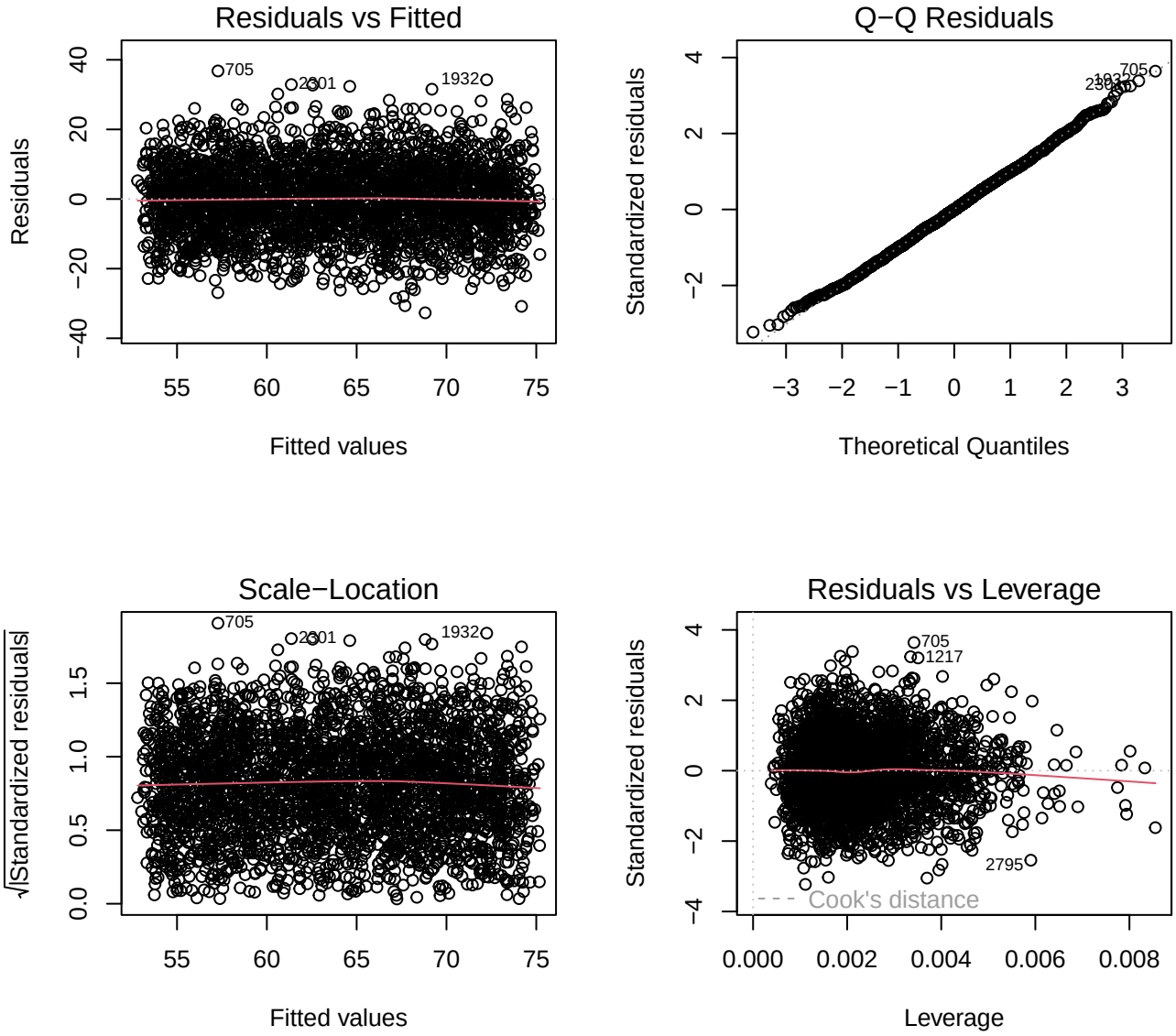
- The regression equation for this linear model can be expressed as:

Cognitive Function = 78.227 - 0.249· *Age* - 0.038· *Hearing Ability* + 2.731· *Vision Sharpness* - 0.141· *Bone Density* + 0.008· *Blood Glucose Level* - 0.006· *Cholesterol Level* + ϵ

- We can see that the baseline Cognitive Function that people who responded to the survey reported when our predictors are zero is 78.227 from a score range of 0 to 100 (Keep in mind that no such state exists).
- We can see that there is a positive coefficient in Vision Sharpness. Which suggests that an increase in Vision Sharpness rating, Cognitive Function rating increases. An increase in Vision Sharpness by one point is related to a 2.731 increase in Cognitive Function rating (holding other predictors constant). However, it should be noted that this effect is not statistically significant (p-value=0.1766).
- The most significant negative factor would have to be Age. As an increase in Age by one year is associated to a 0.249 decrease in Cognitive Function rating (holding other predictors constant).
- An increase of 1 unit in Hearing Ability, Cognitive Function decreases by 0.04 points (holding other predictors constant). This effect is statistically significant (p-value=0.0381).
- An increase of 1 unit in Vision Sharpness, Cognitive Function increases by 2.731 points (holding other predictors constant). This effect is not statistically significant (p-value=0.1766).
- An increase of 1 unit in Bone Density, Cognitive Function decreases by 0.14 points (holding other predictors constant). This effect is not statistically significant (p-value=0.9067).
- An increase of 1 unit in Blood Glucose Level, Cognitive Function increases by 0.008 points (holding other predictors constant). This effect is not statistically significant (p-value=0.4983).
- An increase of 1 unit in Cholesterol Level, Cognitive Function decreases by 0.006 points (holding other predictors constant). This effect is not statistically significant (p-value=0.4834).
- We can see that Age and Hearing Ability are the only predictors where their effects are statistically significant at a significance level of 0.05.
- Although we have only two predictors having a significant effect on Cognitive Function, we can see that our overall model is extremely significant (p-value $< 2 \times 10^{-16}$)
- We can also see that our R^2 is 0.2599. Which implies that approximately 25.99 of the variability in Cognitive Function can be explained by our predictors. However, after adjusting the number of predictors, the model explains about 25.84 of the variability in Cognitive Function. Since our R^2 value is 0.2599, it also suggests that our model is a poor fit.

Dianostic Checking for our full Mulitple Regression Model

Model Assumptions



- Based of the Residual vs Fitted plot, we can see that most of our data entries are scattered around the zero line. We can also see that the points are randomly scattered without a funnel shape or pattern suggesting that our residuals are homoscedastic, and independent. This also indicates that the relationship between our predictors and Cognitive Function are linear. The diagram also shows that there are clearly three outliers 705, 2001, and 1932.
- Based of our Q-Q Residuals plot, we can see that almost all of our points are following the line. Which indicates that our residuals are approximately normal. Which means our normality assumptions of our residuals are met.
- Based on the Residuals vs Leverage plot, we can see that our data is clustered on the left. Which suggests that most of our observations are not extreme predictor values, resulting in low leverage.
- With the assumptions of are residuals are true (meaning there is linearity, homoscedasticity, independence, and Normality of residuals), we do not really need to transform our regression model.

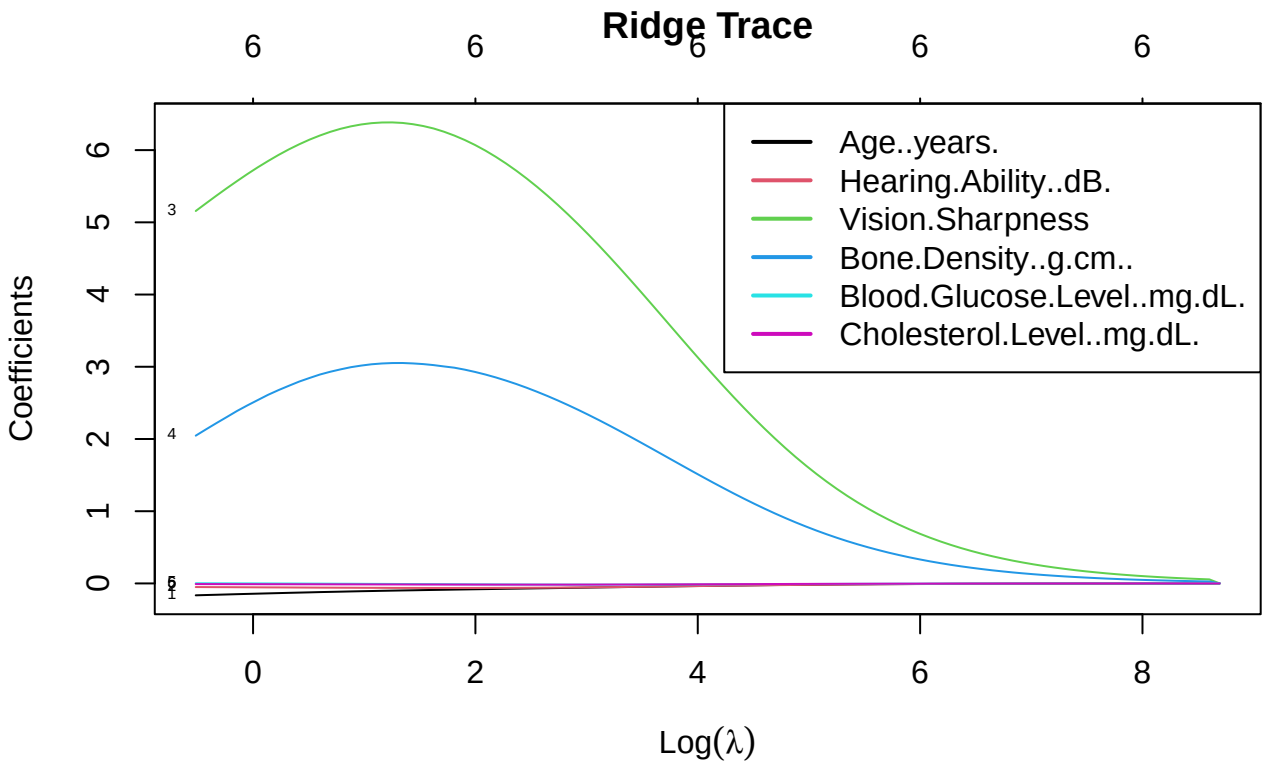
Multicollinearity

```
##                  Age..years.        Hearing.Ability..dB.
##                    13.909221                    2.032736
##             Vision.Sharpness        Bone.Density..g.cm..
##                     5.255452                    8.295757
## Blood.Glucose.Level..mg.dL.   Cholesterol.Level..mg.dL.
##                     1.227182                    1.231819
```

- The VIF results indicate an extremely problematic multicollinearity for Age (13.91) and Bone Density (8.30), suggesting these predictors are highly correlated with others and may destabilize the model. Vision Sharpness (5.26) shows high multicollinearity, while Hearing Ability (2.03), Blood Glucose Level (1.23), and Cholesterol Level (1.23) have low VIFs, indicating having moderate multicollinearity.
- We will need to apply Ridge Regression to deal with Age, Bone Density, and Vision Sharpness as they have high multicollinearity.

Applying Ridge Regression Since we see that the multicollinearity for the predictors Age, Bone Density, and Vision Sharpness, we will be applying ridge regression to see the significance of these predictors and see if we would need to remove them from our regression model.

Ridge Trace Plot



- Based of the Ridge Trace plot, we can see that Vision Sharpness and Bone Density are predictors that have a steep decline. Which suggests that Vision Sharpness and Bone Density has strong influence on Cognitive Function and is affect by regularization. Since they are likely to be affected by regularization, they should be removed from our model.
- We can also see that Age and Hearing Ability are predictors that have a noticeable incline. Which indicate that the predictors are less redundant and carry unique information regarding or model. Which suggests that it has an influence in our data.

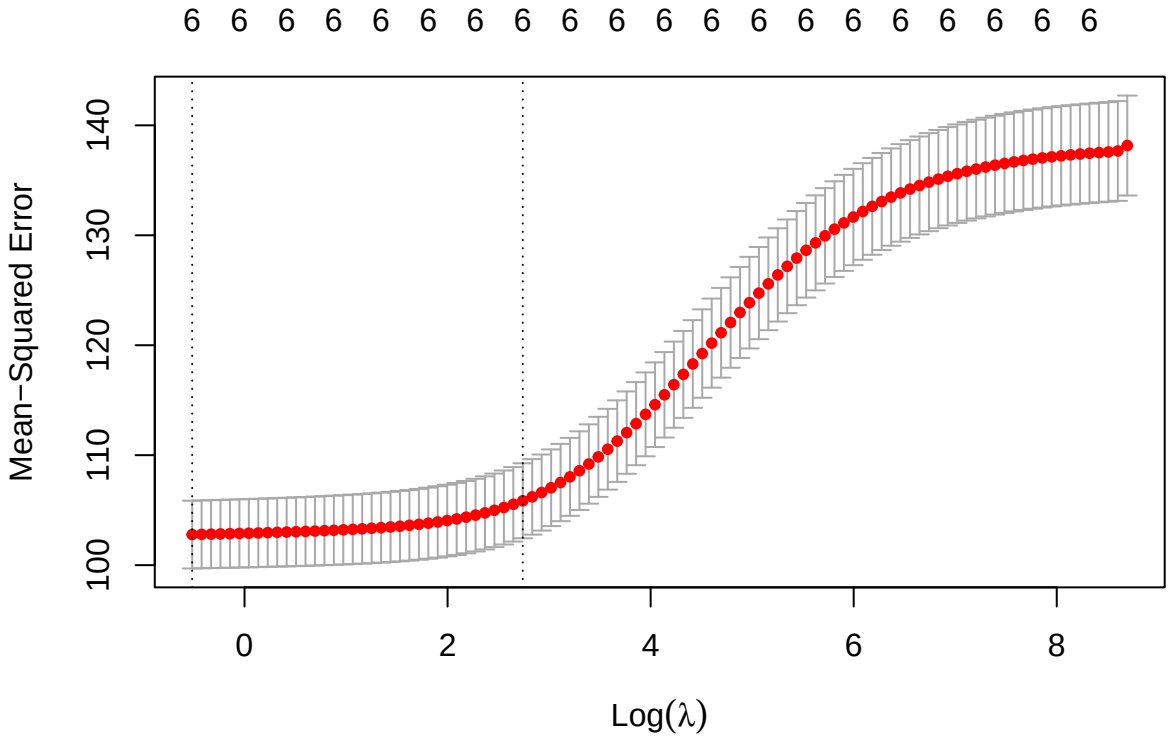
- Cholesterol Level and Blood Glucose Level both have a pretty flat line. Which suggests that Cholesterol Level and Blood Glucose Level have less impact on Cognitive Function. Which can inform us to remove them from our model.

Optimal Lambda

```
## [1] 0.5971598
```

- The optimal value of λ that minimizes the test MSE is 0.5971598.

Cross-Validation Plot



- Based on this graph, the shape suggests that the full model is doing a good job. Which means that the model is effectively capturing the relationships between our predictors and our response variable (Cognitive Function).

Coefficients at Optimal Lambda

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                72.537613634
## Age..years.                -0.164789806
## Hearing.Ability..dB.       -0.050181944
## Vision.Sharpness            5.198104222
## Bone.Density..g.cm..        2.018346202
## Blood.Glucose.Level..mg.dL. 0.002430508
## Cholesterol.Level..mg.dL.  -0.009194040
```

Coefficients of Original Regression Model

```
##
## Call:
## lm(formula = y ~ X)
##
## Coefficients:
##                   (Intercept)                   XAge..years.
##                     78.227262                     -0.249081
##        XHearing.Ability..dB.             XVision.Sharpness
##                     -0.038137                      2.731449
##        XBone.Density..g.cm..  XBlood.Glucose.Level..mg.dL.
##                     -0.140770                      0.007609
##   XCholesterol.Level..mg.dL.
##                     -0.005864
```

- Strictly looking the predictors that had a high VIF values, we can see something interesting. We can observe that there is a significant increase in coefficient with Vision Sharpness and Bone Density. What is more interesting is that the coefficient for Bone Density went from a negative to positive coefficient.

Thoughts - Based on what we have seen from applying a ridge regression, we can see that we should remove many predictors in our model to best explain the variability of our predictors of our model. From the Ridge Trace Plot, it makes us believe that we should remove Vision Sharpness, Bone Density, Blood Glucose Level, and Cholesterol Level. As Vision Sharpness and bone density are variables that are affect by regularization. For Blood Glucose Level and Cholesterol Level, they are seem to have little to no impact on our response variable (Cognitive Function).

R-Squared of our Ridge Regression Model:

```
y_predicted <- predict(ridge_fit, s = best_lambda, newx = X)

#find SST and SSE
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

#find R-Squared
rsq <- 1 - sse/sst
rsq
```

```
## [1] 0.2582572
```

- Looking at the value above, we can see that we have a R^2 value of 0.2582572. Which indicate that the best ridge regression model was able to explain 25.82572% of our variation in the response of our training data.

Variable Selection

To prove or validate our thought from the Ridge Regression we will employ a Stepwise Regression to see which variables we should use to better explain the variability our predictors of our model.

Forward Selection

```
## Start:  AIC=14787.05
## Cognitive.Function ~ 1
##
##                              Df Sum of Sq    RSS   AIC
## + Age..years.                 1    106980 307474 13893
## + Bone.Density..g.cm..        1     93880 320574 14018
```

```
## + Vision.Sharpness             1     90028 324425 14054
## + Hearing.Ability..dB.         1     61409 353045 14308
## + Cholesterol.Level..mg.dL.    1     21964 392490 14626
## + Blood.Glucose.Level..mg.dL.  1     18113 396341 14655
## <none>                                     414454 14787
##
## Step:  AIC=13893.33
## Cognitive.Function ~ Age..years.
##
##                               Df Sum of Sq    RSS   AIC
## + Hearing.Ability..dB.         1    445.18 307029 13891
## <none>                                     307474 13893
## + Vision.Sharpness             1    175.21 307299 13894
## + Cholesterol.Level..mg.dL.    1     56.59 307417 13895
## + Blood.Glucose.Level..mg.dL.  1     38.32 307436 13895
## + Bone.Density..g.cm..         1      0.78 307473 13895
##
## Step:  AIC=13890.99
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB.
##
##                               Df Sum of Sq    RSS   AIC
## <none>                                     307029 13891
## + Vision.Sharpness             1   178.364 306850 13891
## + Cholesterol.Level..mg.dL.    1    46.843 306982 13892
## + Blood.Glucose.Level..mg.dL.  1    40.949 306988 13893
## + Bone.Density..g.cm..         1     0.748 307028 13893
```

```
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years. + Hearing.Ability..dB.,
##     data = health_data)
##
## Coefficients:
##          (Intercept)           Age..years.  Hearing.Ability..dB.
##              80.18191              -0.27135              -0.03829
```

- When starting with no predictors in the model and adding predictors one by one based on contributions to improve the model fit, we can see that Age and Hearing Ability are the predictors that give us the best results.

Backward Elimination

```
## Start:  AIC=13896.29
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB. + Vision.Sharpness +
##     Bone.Density..g.cm.. + Blood.Glucose.Level..mg.dL. + Cholesterol.Level..mg.dL.
##
##                               Df Sum of Sq    RSS   AIC
## - Bone.Density..g.cm..         1       1.4 306754 13894
## - Blood.Glucose.Level..mg.dL.  1      47.0 306799 13895
## - Cholesterol.Level..mg.dL.    1      50.3 306803 13895
## - Vision.Sharpness             1     187.3 306940 13896
## <none>                                     306752 13896
## - Hearing.Ability..dB.         1     441.0 307193 13899
## - Age..years.                  1    5659.9 312412 13949
##
## Step:  AIC=13894.3
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB. + Vision.Sharpness +
##     Blood.Glucose.Level..mg.dL. + Cholesterol.Level..mg.dL.
##
##                               Df Sum of Sq    RSS   AIC
## - Blood.Glucose.Level..mg.dL.  1      46.6 306800 13893
## - Cholesterol.Level..mg.dL.    1      50.6 306804 13893
## - Vision.Sharpness             1     186.8 306941 13894
## <none>                                     306754 13894
## - Hearing.Ability..dB.         1     441.0 307195 13897
## - Age..years.                  1   11490.4 318244 14003
##
## Step:  AIC=13892.76
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB. + Vision.Sharpness +
##     Cholesterol.Level..mg.dL.
##
##                              Df Sum of Sq    RSS   AIC
## - Cholesterol.Level..mg.dL.  1      50.0 306850 13891
## - Vision.Sharpness           1     181.6 306982 13892
## <none>                                    306800 13893
## - Hearing.Ability..dB.       1     438.3 307239 13895
## - Age..years.                1   11543.1 318344 14002
##
## Step:  AIC=13891.24
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB. + Vision.Sharpness
##
##                         Df Sum of Sq    RSS   AIC
## - Vision.Sharpness       1     178.4 307029 13891
## <none>                               306850 13891
## - Hearing.Ability..dB.   1     448.3 307299 13894
## - Age..years.            1   12268.9 319119 14007
##
## Step:  AIC=13890.99
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB.
##
##                         Df Sum of Sq    RSS   AIC
## <none>                               307029 13891
## - Hearing.Ability..dB.   1       445 307474 13893
## - Age..years.            1     46016 353045 14308
```

```
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years. + Hearing.Ability..dB.,
##     data = health_data)
##
## Coefficients:
##          (Intercept)           Age..years.  Hearing.Ability..dB.
##              80.18191              -0.27135              -0.03829
```

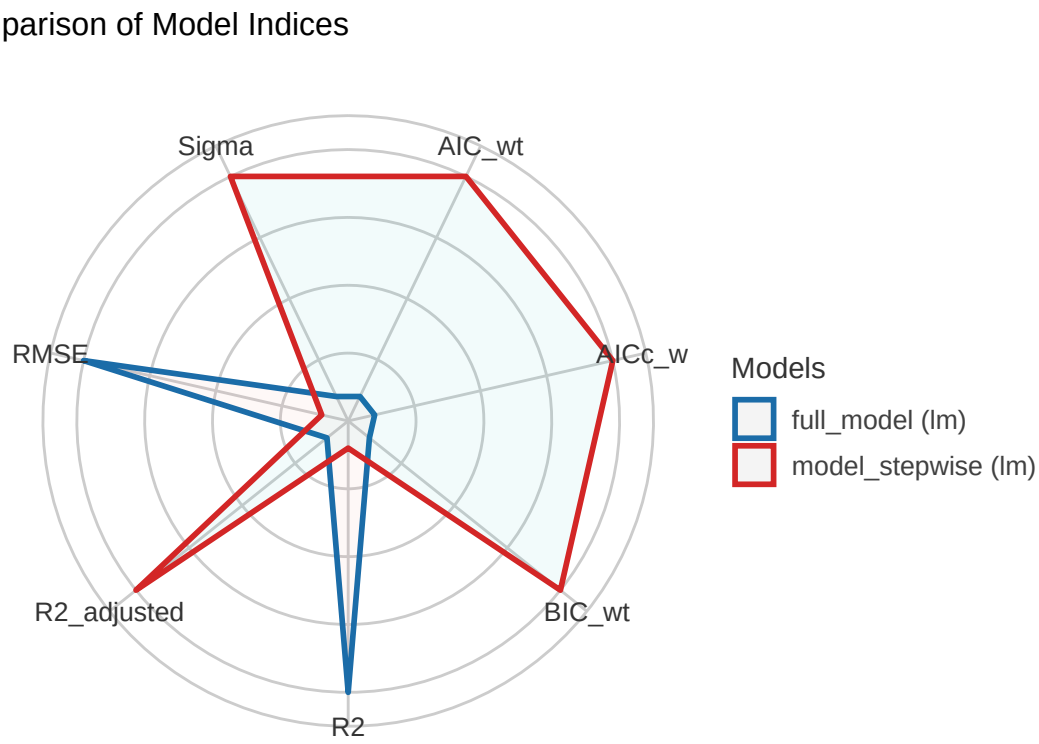
- When starting with all potential predictors in the model and remove predictors one by one, we can see that the predictors that give us the best results are Age and Hearing Ability. Which is the same as when we applied forward selection.

Stepwise Selection

```
## Start:  AIC=14787.05
## Cognitive.Function ~ 1
##
##                               Df Sum of Sq    RSS   AIC
## + Age..years.                  1    106980 307474 13893
## + Bone.Density..g.cm..         1     93880 320574 14018
## + Vision.Sharpness             1     90028 324425 14054
## + Hearing.Ability..dB.         1     61409 353045 14308
## + Cholesterol.Level..mg.dL.    1     21964 392490 14626
## + Blood.Glucose.Level..mg.dL.  1     18113 396341 14655
## <none>                                     414454 14787
##
## Step:  AIC=13893.33
## Cognitive.Function ~ Age..years.
##
##                               Df Sum of Sq    RSS   AIC
## + Hearing.Ability..dB.         1       445 307029 13891
## <none>                                     307474 13893
## + Vision.Sharpness             1       175 307299 13894
## + Cholesterol.Level..mg.dL.    1        57 307417 13895
## + Blood.Glucose.Level..mg.dL.  1        38 307436 13895
## + Bone.Density..g.cm..         1         1 307473 13895
## - Age..years.                  1    106980 414454 14787
##
## Step:  AIC=13890.99
## Cognitive.Function ~ Age..years. + Hearing.Ability..dB.
##
##                               Df Sum of Sq    RSS   AIC
## <none>                                     307029 13891
## + Vision.Sharpness             1       178 306850 13891
## + Cholesterol.Level..mg.dL.    1        47 306982 13892
## + Blood.Glucose.Level..mg.dL.  1        41 306988 13893
## + Bone.Density..g.cm..         1         1 307028 13893
## - Hearing.Ability..dB.         1       445 307474 13893
## - Age..years.                  1     46016 353045 14308
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years. + Hearing.Ability..dB.,
##     data = health_data)
```

```
##
## Coefficients:
##          (Intercept)           Age..years.  Hearing.Ability..dB.
##              80.18191              -0.27135              -0.03829
```

- When combining both forward selection and backward elimination, we can see that the predictors that give the best fit are Age and Hearing Ability. Since all of them are the same, we will be using the stepwise model analysis.



- We can see that based on this diagram, that the Stepwise model has a larger polygon. Which suggests that it will be a better model.

```
##
## Call:
## lm(formula = Cognitive.Function ~ Age..years. + Hearing.Ability..dB.,
##     data = health_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.728  -6.839  -0.083   6.876  37.244
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          80.18191    0.63747 125.781   <2e-16 ***
## Age..years.          -0.27135    0.01280 -21.194   <2e-16 ***
## Hearing.Ability..dB. -0.03829    0.01837  -2.085   0.0372 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.12 on 2997 degrees of freedom
## Multiple R-squared:  0.2592, Adjusted R-squared:  0.2587
## F-statistic: 524.3 on 2 and 2997 DF,  p-value: < 2.2e-16
```

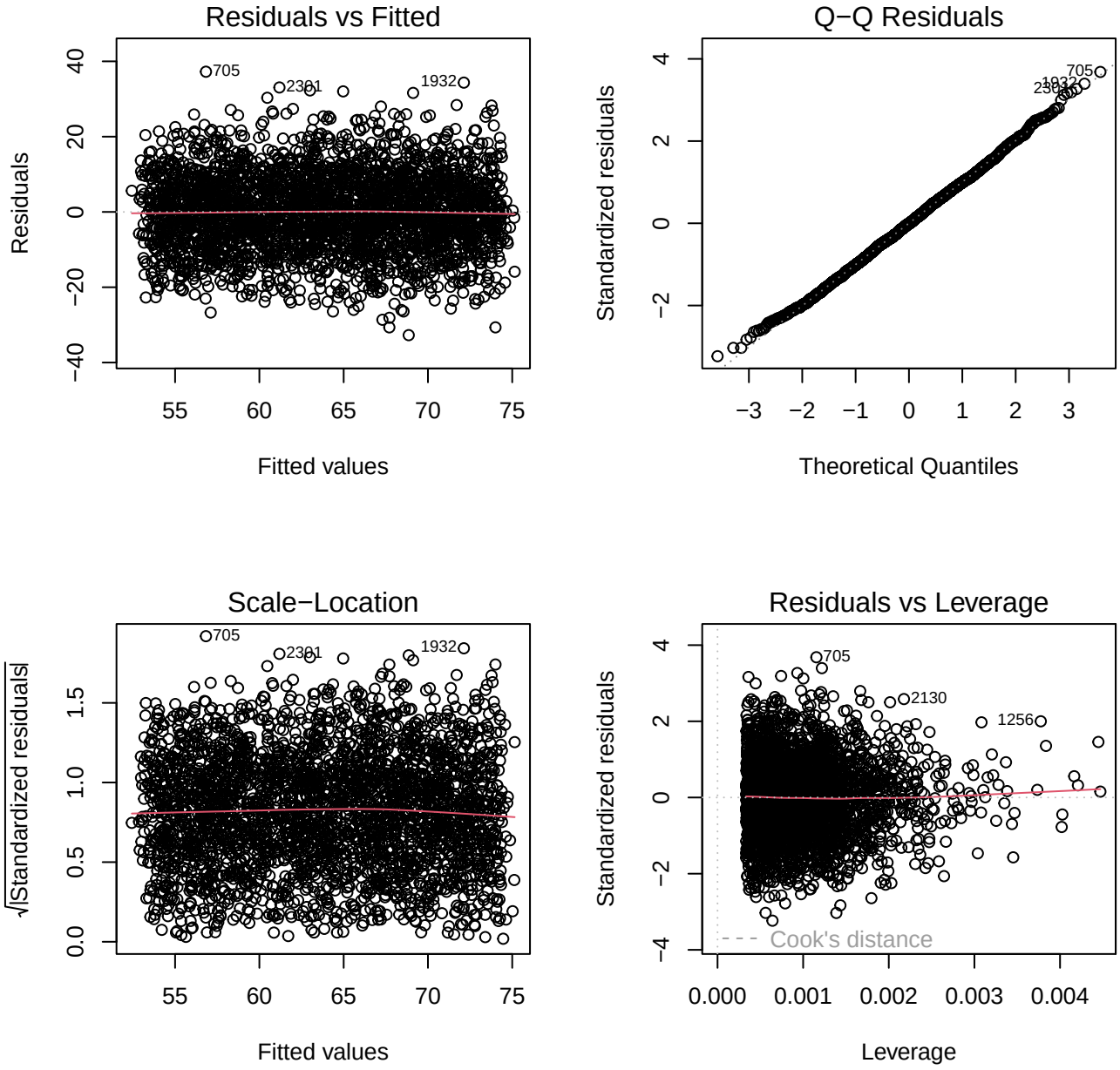
- The regression equation for this linear model can be expressed as:

$$Cognitive\ Function = 80.182 - 0.271 * Age - 0.038 * Hearing\ Ability + \epsilon$$

- We can see that the baseline Cognitive Function that people who responded to the survey reported when our Age and Hearing Ability predictor is zero is 80.182 from a score range of 0 to 100 (Keep in mind that no such state exists).
- Based on the model, we can see that as age increases by one year cognitive function decreases by 0.271. This effect is highly statistically significant (p-value $<2 \times 10^{-16}$).
- Based on the model, we can see that as age increases by one year cognitive function decreases by 0.038. This effect is highly statistically significant (p-value =0.0372).
- We can also see that our overall model is extremely significant (p-value $<2 \times 10^{-16}$)
- We can also see that our R^2 is 0.2592. Which implies that approximately 25.92 of the variability in Cognitive Function can be explained by Age. However, after adjusting the number of ages, the model explains about 25.87 of the variability in Cognitive Function. Since our R^2 value is 0.2592, it also suggests that our model is a poor fit.

Diagnostic Check of the Best Model

Model Assumptions



- Based of the Residual vs Fitted plot, we can see that most of our data entries are scattered around the zero line. We can also see that the points are randomly scattered without a funnel shape or pattern suggesting that our residuals are homoscedastic, and independent. This also indicates that the relationship between our predictors and Cognitive Function are linear. The diagram also shows that there are clearly three outliers at observation 705, 2001, and 1932.
- Based of our Q-Q Residuals plot, we can see that almost all of our points are following the line. Which indicates that our residuals are approximately normal. Which means our normality assumptions of our residuals are met.
- Based on the Residuals vs Leverage plot, we can see that our data is clustered on the left. Which suggests that most of our observations are not extreme predictor values, resulting in low leverage. We

can notice that our points are even more spread to the left compared to our full model.

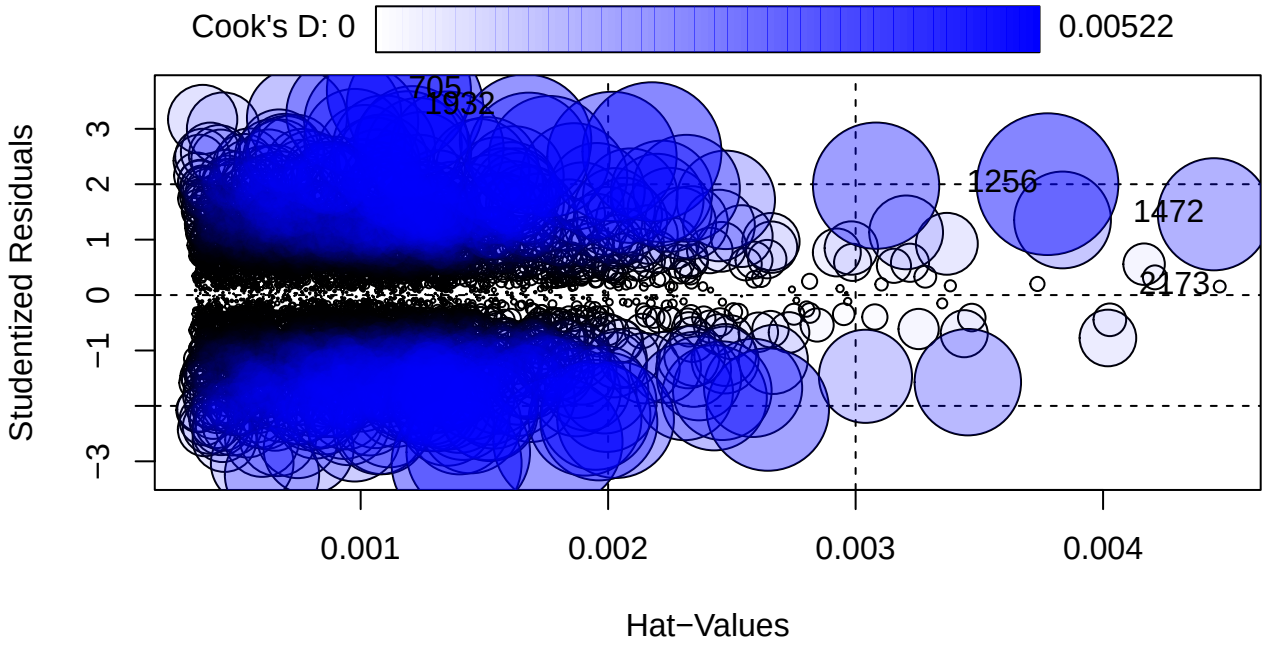
- With the assumptions of are residuals are true (meaning there is linearity, homoscedasticity, independence, and Normality of residuals), we do not really need to transform our regression model.

Multicollinearity

```
##          Age..years. Hearing.Ability..dB.
##             2.030362             2.030362
```

- We can see that unlike the full model, our predictors do not have problematic mulitcollinearity to worry about. Whcih suggests that there are no predictors that we would need to shrink or remove from our model.

Influential Points and Outliers **Influential Points**




```
##        StudRes         Hat        CookD
## 705  3.6895882 0.001154556 5.223086e-03
## 1256 2.0012965 0.003775278 5.054264e-03
## 1472 1.4579930 0.004447863 3.164564e-03
## 1932 3.4016959 0.001219073 4.691372e-03
## 2173 0.1540127 0.004471295 3.552328e-05
```

- We can see that there are 5 points $(705, 1256, 1472, 1932, and\ 2173)$ are identified as influential points.

Outliers

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 705 3.689588         0.00022861      0.68584
```

- Point 705 is the only point that is identified as an outlier from our model.

Conclusion

Circling back to our first research question and hypothesis that we are testing. Are there any significant associations between Physical Health metrics and Cognitive Function? Based on the intensive regression analysis that we have done, we can conclude that there is sufficient evidence to claim that there are only two significant **Physical** Health metrics that have a significant negative impact on Cognitive Function. These two predictors are *Age* and *Hearing Ability*. In which as one's age and hearing ability rating increases one's cognitive function rating decreases. It is interesting to find that as one's hearing ability rating increases their cognitive function rating decreases. This might be because when one's hearing ability is great they do not need to use much of their brain power to listen to conversations or lectures. Which suggests a decrease in cognitive function in our research. This research is extremely interesting mainly because one would think that other predictors such as *Vision Sharpness* and *Blood Glucose Levels* do not have significant associations to Cognitive Function. As Vision Sharpness is a common barrier to knowledge. Which can hinder one's ability to think. With Blood Glucose Levels, it is mainly because glucose serves as the human brain's primary source of energy. Hence its impact/importance in one's cognitive function. We should note that when we try to uncover the best model for understanding our response variable, we discovered that the predictors that best explain *Cognitive Function* are *Age* and *Hearing Ability* based on **Stepwise Regression**. Overall, we can say that certain physical health metrics have a significant impact on *Cognitive Function*. However, there are still relevant physical health metrics that we can collect and test to see their significance on *Cognitive Function*.

Age Lifestyle Research (Research Question 2)

Simple Linear Model

```
##
## Call:
## lm(formula = Age..years. ~ Smoking.Status, data = health_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.079 -18.045  -1.034  17.921  40.966
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            57.0794     0.7170  79.608   <2e-16 ***
## Smoking.StatusFormer   -1.2708     0.9270  -1.371    0.171
## Smoking.StatusNever    -9.0453     0.9547  -9.475   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.19 on 2997 degrees of freedom
## Multiple R-squared:  0.03712,    Adjusted R-squared:  0.03648
## F-statistic: 57.77 on 2 and 2997 DF,  p-value: < 2.2e-16
```

- The regression equation for this linear model can be expressed as:

$$Age = 57.0794 - 1.2708 * Smoking_{FormerSmoker} - 9.0453 * Smoking_{NeverSmoked} + \epsilon$$

- The baseline age for respondents to the survey who identified as current smokers is 57.0794 years old. Those who identified as occasional alcohol drinkers had an expected age of 1.2708 years younger, and those who identified as having never smoked had an expected age of 9.0453 years younger.
- The model has a p-value of $< 2 \times 10^{-16}$, suggesting that there is a statistically significant relationship between age and smoking status.
- We can also see that our R^2 is 0.03712. Which implies that approximately 3.712 of the variability in Age can be explained by our predictor. However, after adjusting the number of predictors, the model explains about 3.648 of the variability in Cognitive Function. Since our R^2 value is 0.03712, it also suggests that our model is a poor fit.

Multiple Linear Regression

```
##
## Call:
## lm(formula = Age..years. ~ 0 + Alcohol.Consumption + Physical.Activity.Level +
##     Smoking.Status, data = health_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.521 -18.149  -0.714  17.678  42.289
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## Alcohol.ConsumptionFrequent      55.7385     1.1597  48.062   <2e-16 ***
## Alcohol.ConsumptionNone          56.7965     1.0885  52.178   <2e-16 ***
## Alcohol.ConsumptionOccasional    57.5972     1.0926  52.714   <2e-16 ***
## Physical.Activity.LevelLow        0.7246     1.0211   0.710    0.478
## Physical.Activity.LevelModerate   0.1185     0.9386   0.126    0.900
## Smoking.StatusFormer             -1.3144     0.9275  -1.417    0.157
## Smoking.StatusNever              -9.0272     0.9549  -9.453   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.19 on 2993 degrees of freedom
## Multiple R-squared:  0.8762, Adjusted R-squared:  0.8759
## F-statistic:  3025 on 7 and 2993 DF,  p-value: < 2.2e-16
```

- The regression equation for this linear model can be expressed as:

$\textbf{Age} = 55.7385 \cdot Alcohol_{ConsumptionFrequent} + 56.7965 \cdot Alcohol_{ConsumptionNone} + 57.5972 \cdot Alcohol_{ConsumptionOccasional} + 0.7246 \cdot Physical\ Activity\ Level_{Low} + 0.1185 \cdot Physical\ Activity\ Level_{Moderate}$ - $1.3144 \cdot Smoking_{FormerSmoker}$ - $9.0272 \cdot Smoking_{NeverSmoked} + \epsilon$

- Let seems that our model does not have a baseline value for *Age*.
- We can see that *Alcohol Consumption* has an extremely high coefficient. Which would make sense that they have strong statistically significance on our response variable.
- The model has a p-value of $< 2 \times 10^{-16}$, suggesting that our model is significant. Individually, the three alcohol consumption levels and having never smoked are the only predictors that are extremely statistically significant (all p-value$< 2 \times 10^{-16}$).
- We can also see that our R^2 is 0.8762. Which implies that approximately 87.62 of the variability in Age can be explained by our predictor. However, after adjusting the number of predictors, the model explains about 87.59 of the variability in Cognitive Function. Since our R^2 value is 0.8762, it also suggests that our model is a good fit.

Variable Selection

Forward Selection

```
## Start:  AIC=18143.91
## Age..years. ~ 1
##
##                           Df Sum of Sq     RSS   AIC
## + Smoking.Status           2     47107 1221807 18034
## + Alcohol.Consumption      2      1970 1266944 18143
## <none>                                 1268913 18144
```

```
## + Physical.Activity.Level  2       341 1268572 18147
##
## Step:  AIC=18034.42
## Age..years. ~ Smoking.Status
##
##                           Df Sum of Sq     RSS   AIC
## <none>                                  1221807 18034
## + Alcohol.Consumption      2    1530.7 1220276 18035
## + Physical.Activity.Level  2     294.5 1221512 18038

##
## Call:
## lm(formula = Age..years. ~ Smoking.Status, data = health_data)
##
## Coefficients:
##          (Intercept)  Smoking.StatusFormer   Smoking.StatusNever
##               57.079                -1.271                -9.045
```

- When starting with no predictors in the model and adding predictors one by one based on contributions to improve the model fit, we can see that Smoking status is the predictor that give us the best results.

Backward Elimination

```
## Start:  AIC=18038
## Age..years. ~ 0 + Alcohol.Consumption + Physical.Activity.Level +
##     Smoking.Status
##
##                           Df Sum of Sq     RSS   AIC
## - Physical.Activity.Level  2       269 1220276 18035
## <none>                                  1220007 18038
## - Smoking.Status           2     46629 1266636 18146
## - Alcohol.Consumption      3   1379515 2599522 20301
##
## Step:  AIC=18034.66
## Age..years. ~ Alcohol.Consumption + Smoking.Status - 1
##
##                       Df Sum of Sq     RSS   AIC
## <none>                               1220276 18035
## - Smoking.Status       2     46668 1266944 18143
## - Alcohol.Consumption  3   2585175 3805451 21441

##
## Call:
## lm(formula = Age..years. ~ Alcohol.Consumption + Smoking.Status -
##     1, data = health_data)
##
## Coefficients:
##   Alcohol.ConsumptionFrequent        Alcohol.ConsumptionNone
##                        56.003                         57.057
## Alcohol.ConsumptionOccasional           Smoking.StatusFormer
##                        57.877                         -1.303
##           Smoking.StatusNever
##                        -9.025
```

- When starting with all potential predictors in the model and remove predictors one by one, we can see that the predictors that give us the best results is Smoking Status, which matches our results from

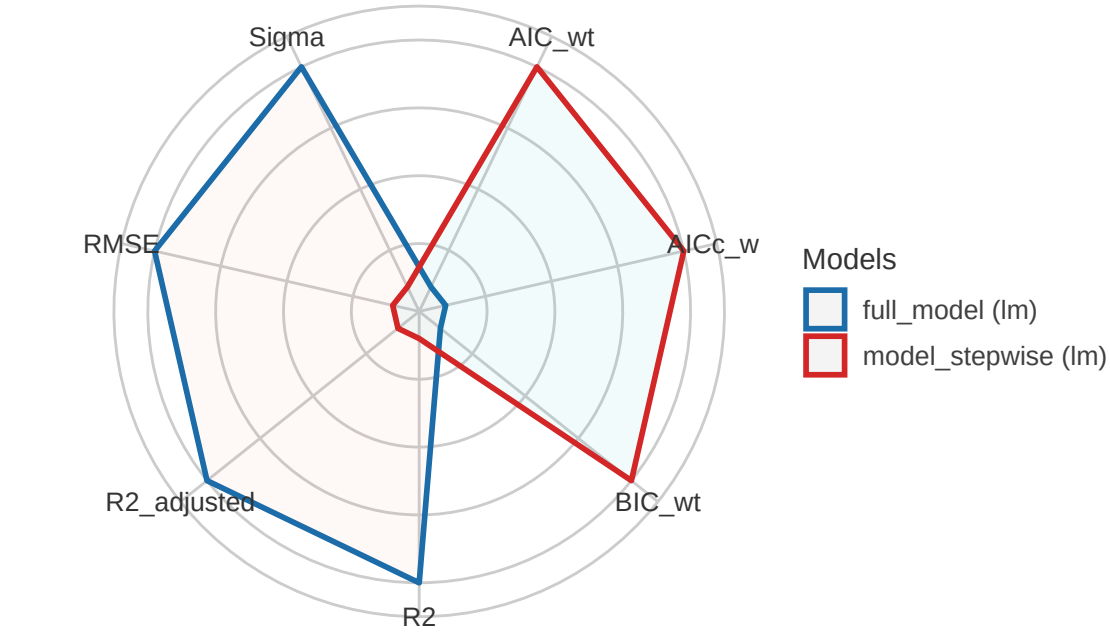
when we applied forward selection.

Stepwise Selection

```
## Start:  AIC=18143.91
## Age..years. ~ 1
##
##                            Df Sum of Sq     RSS   AIC
## + Smoking.Status            2     47107 1221807 18034
## + Alcohol.Consumption       2      1970 1266944 18143
## <none>                                  1268913 18144
## + Physical.Activity.Level   2       341 1268572 18147
##
## Step:  AIC=18034.42
## Age..years. ~ Smoking.Status
##
##                            Df Sum of Sq     RSS   AIC
## <none>                                  1221807 18034
## + Alcohol.Consumption       2      1531 1220276 18035
## + Physical.Activity.Level   2       294 1221512 18038
## - Smoking.Status            2     47107 1268913 18144
```

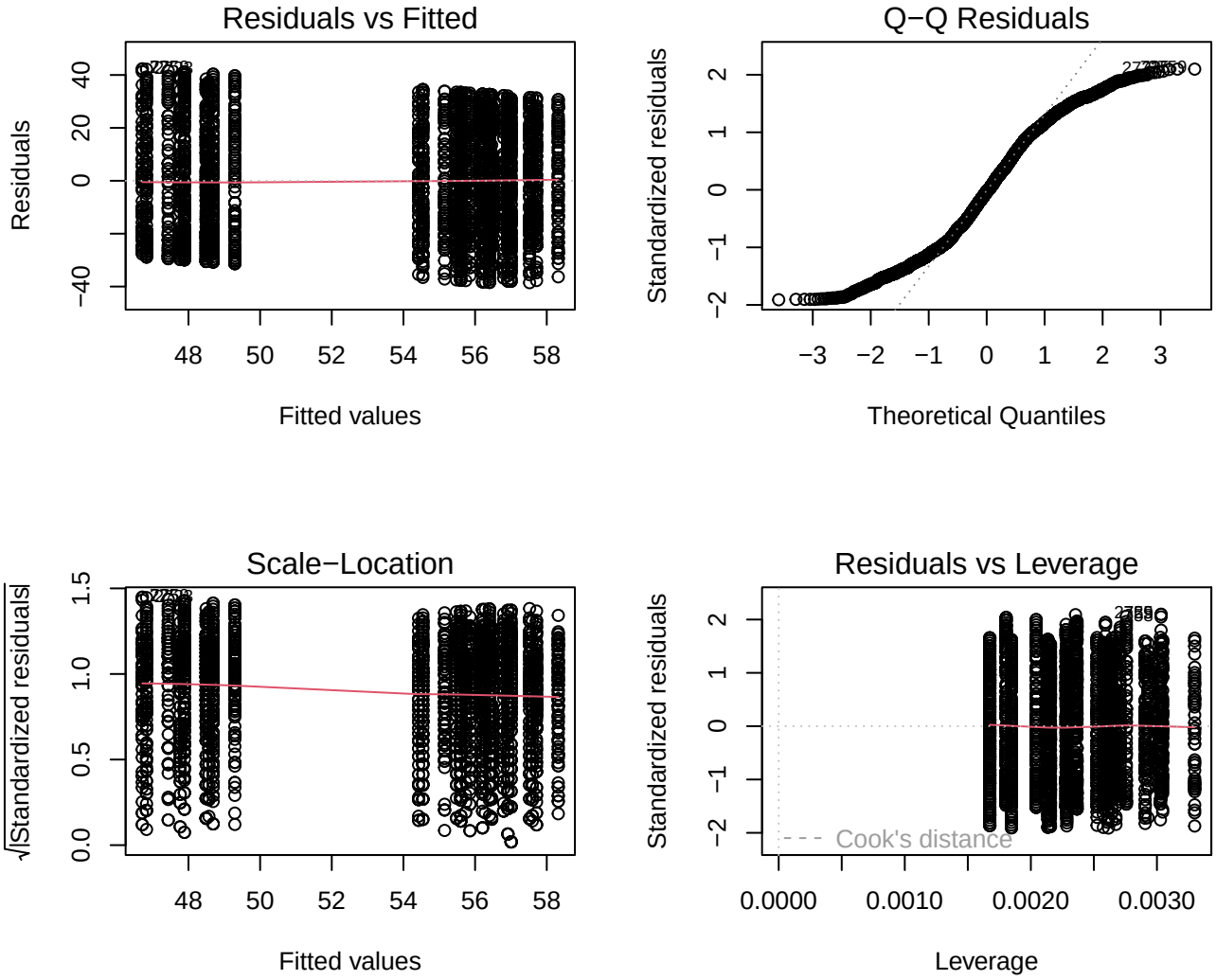
```
##
## Call:
## lm(formula = Age..years. ~ Smoking.Status, data = health_data)
##
## Coefficients:
##          (Intercept)  Smoking.StatusFormer   Smoking.StatusNever
##               57.079                -1.271                -9.045
```

- When combining both forward selection and backward elimination, we can see that the predictor that gives the best fit is once again Smoking Status. Since all of them are the same, we will be using the stepwise model analysis.



- We can see that based on this diagram that the full model has a larger polygon. Which suggests that it will be a better model. We do not need to remove any variables from our model.

Diagnostic Checking



- Based of the Residual vs Fitted plot, we can see that most of our data entries are scattered around the zero line. We can also see that the points are randomly scattered without a funnel shape or pattern suggesting that our residuals are homoscedastic, and independent. This also indicates that the relationship between our predictors and Cognitive Function are linear. The diagram also shows that there are a couple of outliers in the top left.
- Based of our Q-Q Residuals plot, we can see that there is a clear trend to the line that does not match the line, indicating that our residuals are approximately not normal. Rather, the data appears to follow a cumulative distribution.
- Based on our Residual vs Leverage plot, the data plots are all scattered on the right side of the plot. Which suggests that our points have high leverage. Points with high leverage indicate that our predictor values are far from the mean. Which can have a strong influence on the regression goodness of fit.
- Since the normality assumption has been violated, we need to transform our regression model.

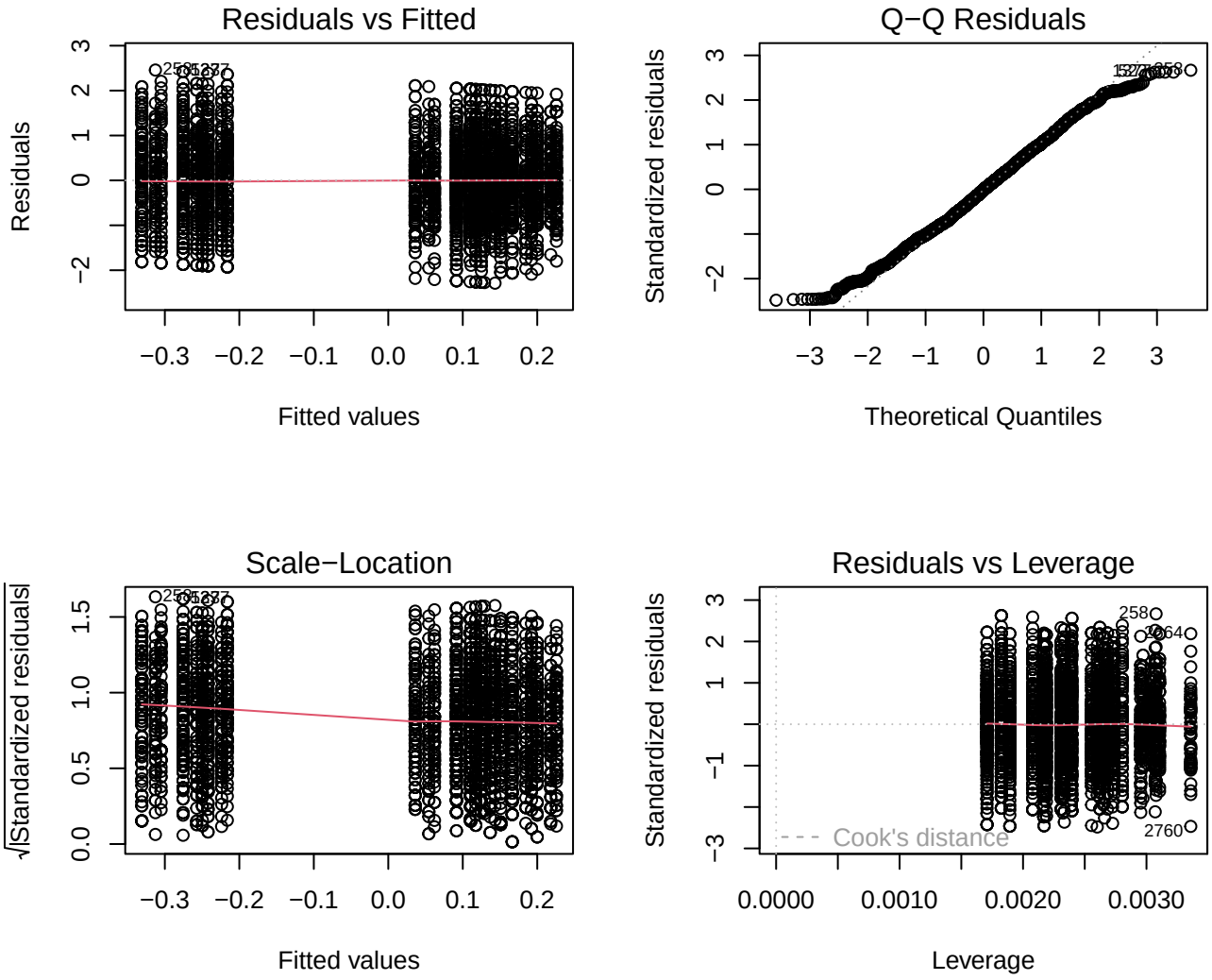
Transforming Using Empirical Cumulative Distribution Function (ECDF) Since our model Q-Q Residual plot does not meet the assumption of normality for our residuals, we will utilize a ECDF transformation to solve this issue.

```
##
##  Shapiro-Wilk normality test
##
## data:  mlr$residuals
## W = 0.96367, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_model_transformed$residuals
## W = 0.99524, p-value = 3.854e-08
```

- Shown by the Shapiro-Wilk test, this transformation has taken our model's normality from 96.37% to 99.52%.

Diagnostic Check on Newly Transformed Model



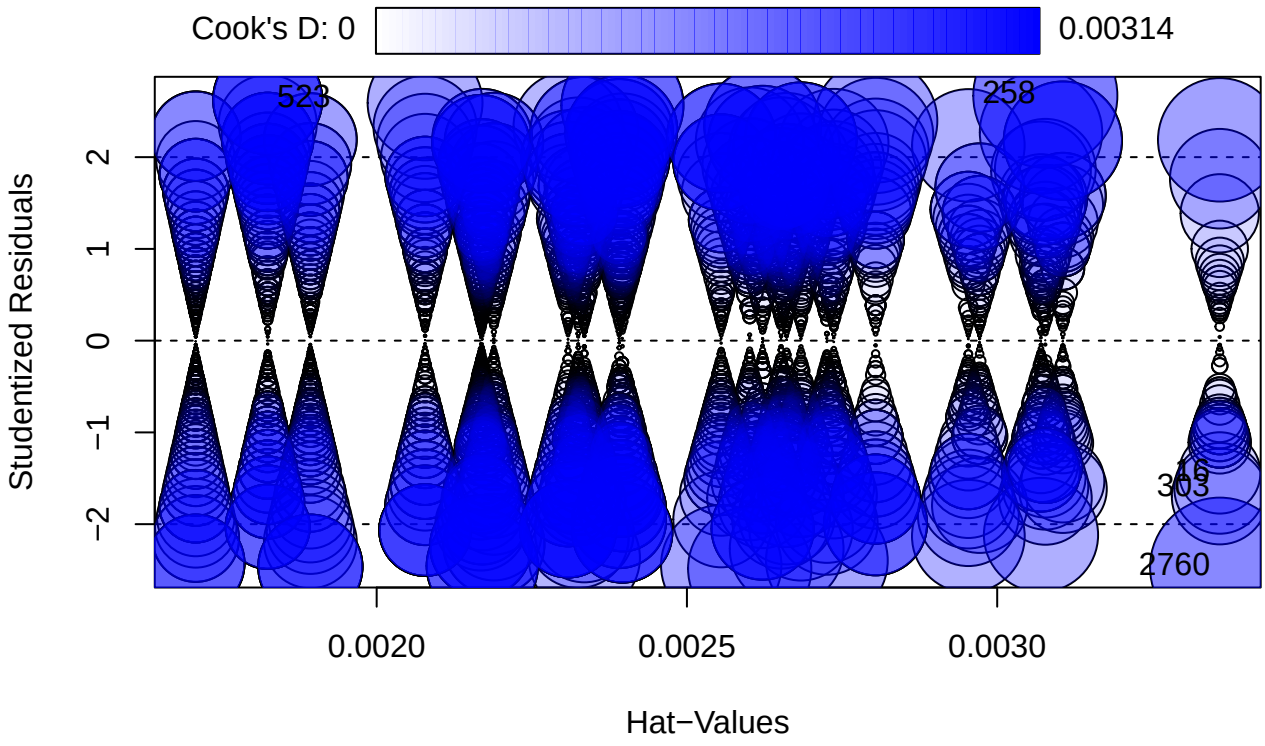
- Based on the matrix plot above we can clearly see that our normality is more stabilized compared to our original model. As we can see that our points are following the line path better.

Multicollinearity

```
##                              GVIF Df GVIF^(1/(2*Df))
## Alcohol.Consumption     1.004138  2        1.001033
## Physical.Activity.Level 1.003110  2        1.000777
## Smoking.Status          1.001982  2        1.000495
```

- All three of our predictor variables have a low VIF, indicating that they are likely not highly correlated with one another and will not affect our model with multicollinearity.

Influential Points




```
##         StudRes         Hat       CookD
## 16   -1.441636 0.003358595 0.001000169
## 258   2.669533 0.003077541 0.003136264
## 303  -1.609549 0.003358595 0.001246506
## 523   2.627375 0.001824426 0.001798854
## 2760 -2.470131 0.003358595 0.002932303
```

- There are 5 influential points at observations 16, 258, 303, 523, and 2760.

Outliers

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 725 2.098957           0.035904           NA
```

- There is only one outlier identified in our dataset at observation 725

Conclusion

The aim of our model was to tackle the question of whether our lifestyle choices (physical activity, smoking, alcohol consumption, etc) has a significant effect on age prediction. Based on the results of our regression model, we can conclude that there is a meaningful connection between the two, with an extra emphasis on the effects of one's smoking status as a predictor. However, there are still aspects to improve upon. While our Empirical CDF transformation was able to achieve normality within our model, it also had the negative effect of dramatically decreasing our R squared value, adding more unexplained variability. In order to improve upon this, we may want to gather more relevant variables, such as hours of sleep per night or dieting habits.