David Jackson

Rick Brown

MATH-3150-01

07 December 2024

Final project

Logistic regression to predict heart disease was the data set I used it comes from the World Health Organization and has over 4,000 records and 15 variables. We had two demographic variables Sex: male(1) or female(0) and age. There were two Behavioral variables whether or not you are a current smoker and how many cigarettes per day you have. There were four medical history variables: BP Meds: whether or not the patient was on blood pressure medication, Prevalent Stroke: whether or not the patient had previously had a stroke, Prevalent Hyp: whether or not the patient was hypertensive, and diabetes: whether or not the patient had diabetes. There were four current medical history variables: Tot Chol: total cholesterol level, Sys BP: systolic blood pressure, Dia BP: diastolic blood pressure, BMI: Body Mass Index, heart rate, and glucose level. The predictor variable was 10-year risk of coronary heart disease CHD that is 1 if they are at risk and 0 if they are not at risk.

For model selection, I used the step function in R to run backward, forward, and both directions, which is a stepwise algorithm that filters variables in the model associated with the AIC statistic. The backward and both directions gave the same model, and the forward direction gave a model with a few more variables in it, so we went with the simpler model for analysis.

The variables in our model were male, age, cigsPerDay, prevalentStroke, prevalentHyp, totChol, sysBP, and glucose.

$$TenYearCHD = -8.739521 + 0.553152male + 0.065337age + 0.019574cigsPerDay + 0.751412prevalentStroke + 0.226231prevalentHyp + 0.002248totChol + 0.014219sysBP + 0.007314glucose$$

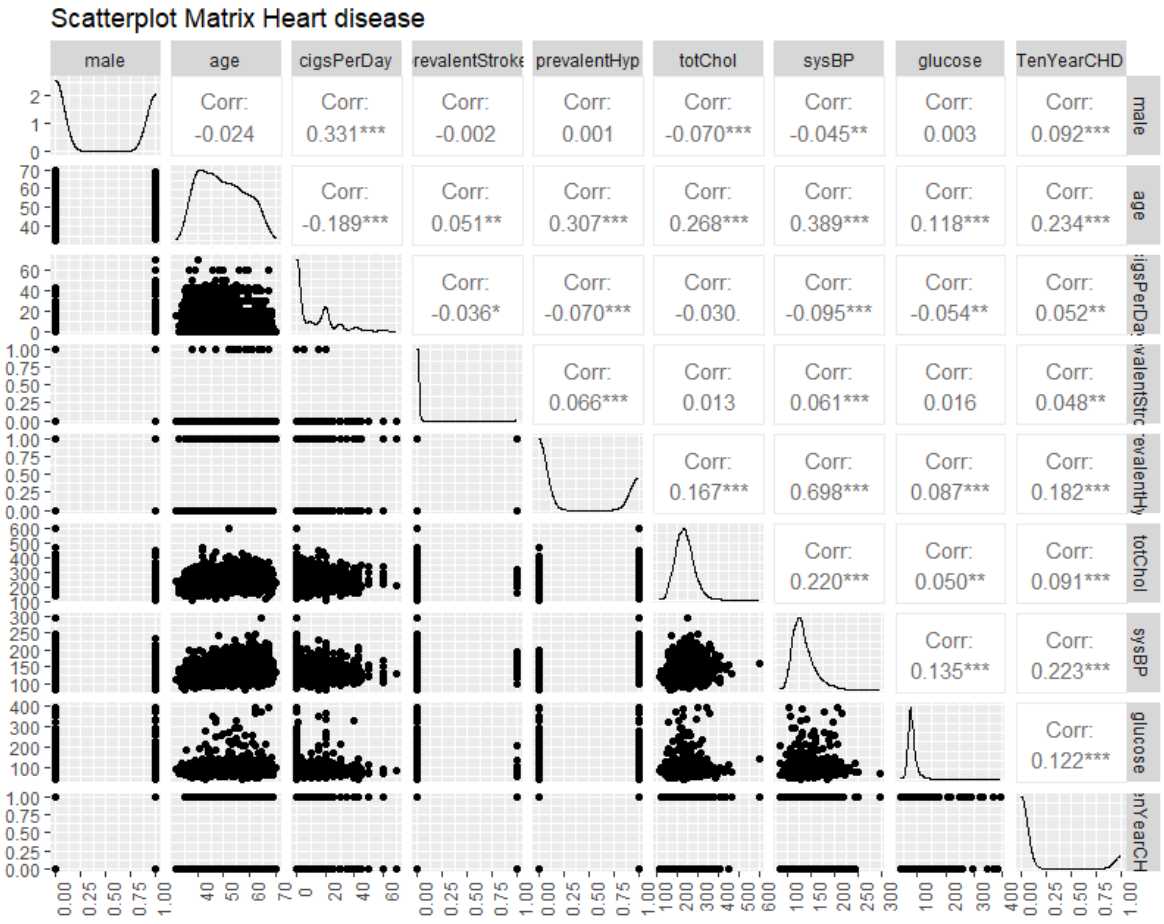
After getting the slopes of our model using the summary function we can interpret some of the slopes as. For age variable: as age increases by 1 the predicted odds that you will have ten years of CHD is increased by $e^{0.065337} = 1.067519$ holding all other variables constant. For the prevalentHyp variable: if you have prevalentHyp the predicted odds that you have ten-year CHD is increased by $e^{0.751412} = 1.253865$ holding all other variables constant. Also, note that in the summary we found the variance inflation factors and they are quite high with this model so interpreting the slopes may not have much meaning because we may not be able to hold the other variables constant but the main thing that we want with this heart disease data set is to be able to predict whether or not someone is at risk for 10-year coronary heart disease so we may not need a to interpret the slope anyways. The next thing we did was run an ANOVA test and get the following table.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TenYearCHD

Terms added sequentially (first to last)

               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)

NULL                           3655     3120.5

male            1   30.567      3654     3090.0 3.225e-08 ***
```

```
age               1  206.285      3653     2883.7 < 2.2e-16 ***
cigsPerDay        1   19.632      3652     2864.0 9.387e-06 ***
prevalentStroke   1    4.137      3651     2859.9   0.04196 *
prevalentHyp      1   48.999      3650     2810.9 2.561e-12 ***
totChol           1    6.022      3649     2804.9   0.01412 *
sysBP             1   28.514      3648     2776.4 9.301e-08 ***
glucose           1   19.176      3647     2757.2 1.192e-05 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

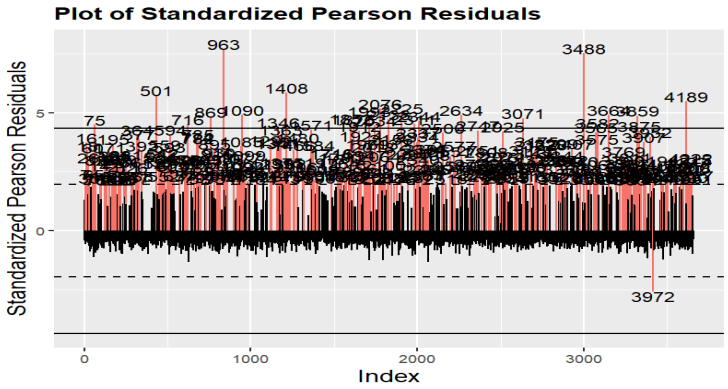
The results from the Analysis of Deviance Table show how each explanatory contributes to reducing the model's deviance when added sequentially. Male, age, cigsPerDay, prevalentHyp, sysBP, and glucose significantly reduce deviance with very small p-values whereas prevalent stroke and total cholesterol are not as significant with larger p-values.

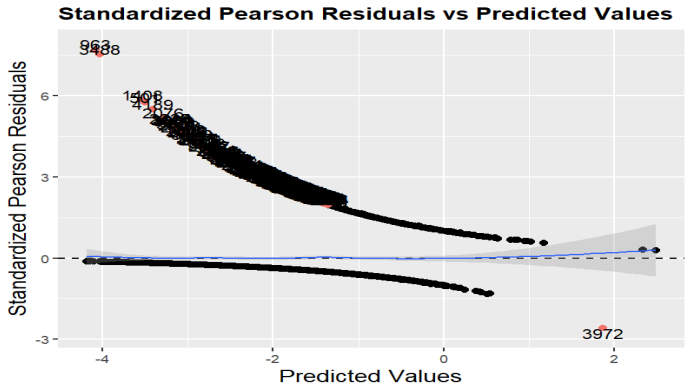
The correlation matrix

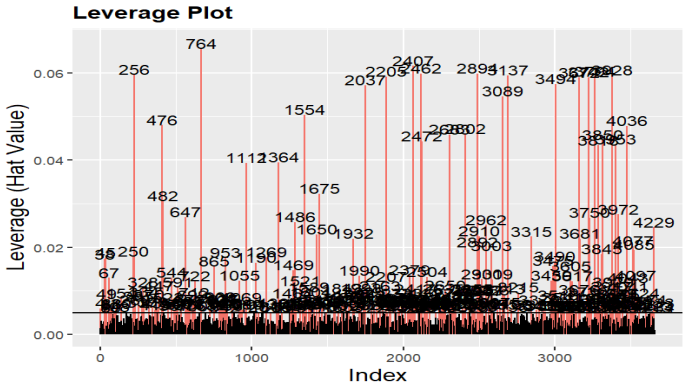


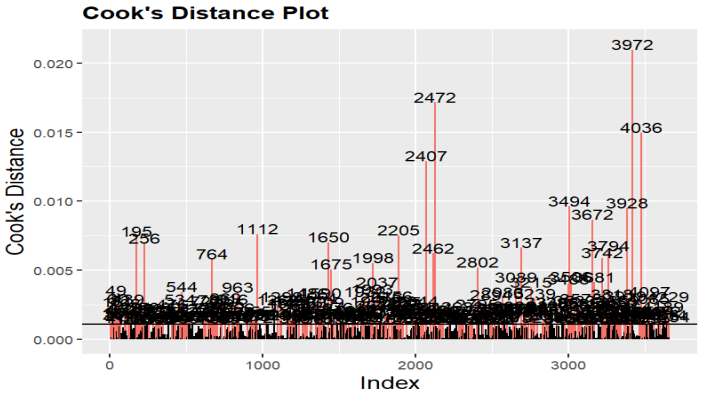
As we can see from the correlation Matrix systolic blood pressure and prevention hypertension are highly correlated with each other and also had high VIF scores so that could be causing some multicollinearity.

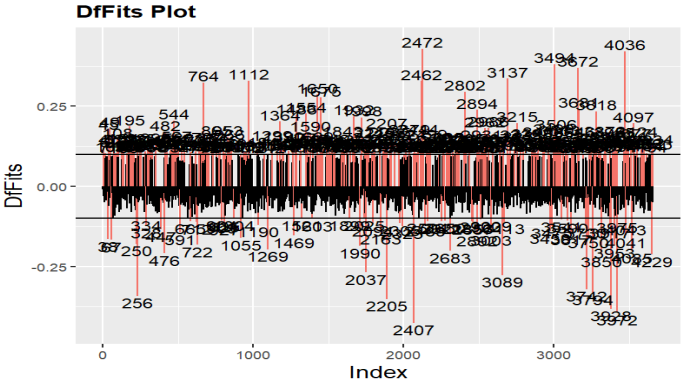
The next thing we checked was the logistic plots to see whether or not our model assumptions were met.

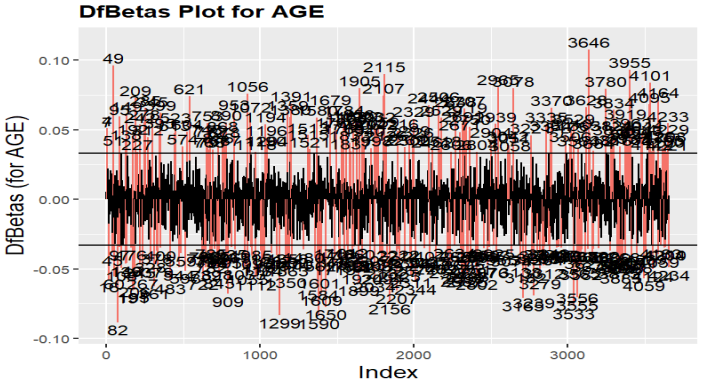






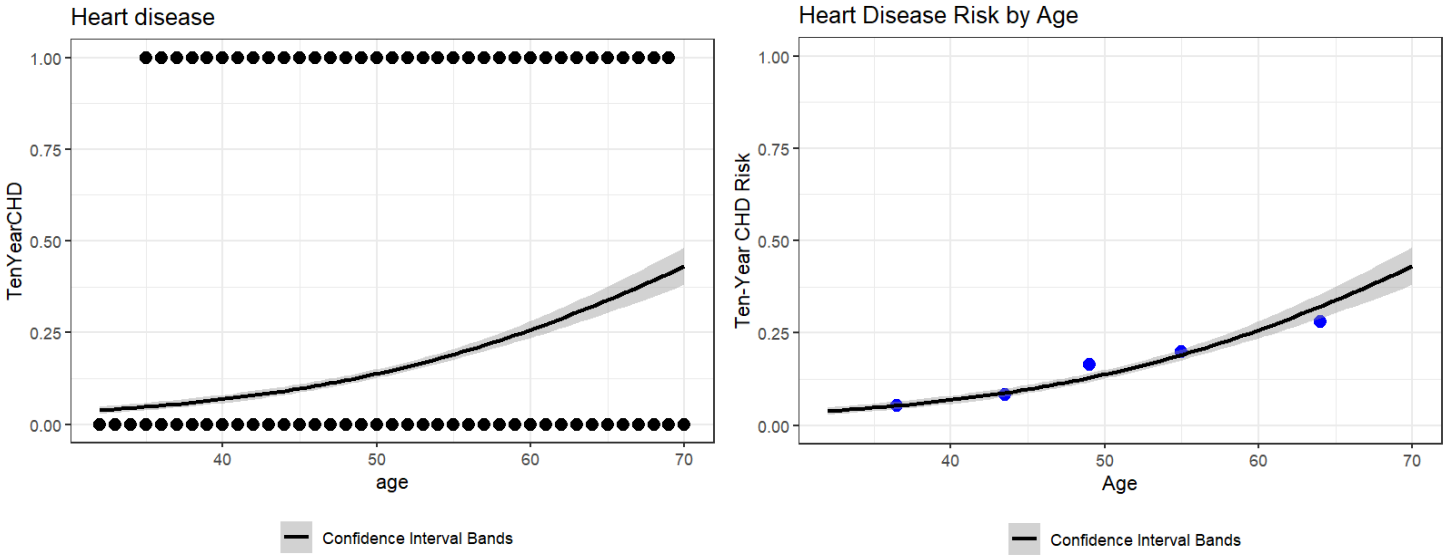






As we can see from the plots nothing seems too out of the ordinary and most importantly in the residual plot the smoothed conditional means line the blue one is mostly horizontal and at zero. The 95 confidence interval always captures zero in it as well which suggests that our model assumptions are met that is the variables are independent of one another and the relationship between them is linear.

The graph on the left below shows the logistic regression age variable in the model, and the graph on the right shows the logistic regression using the age variable broken into age groups ("32-41", "42-46", "47-52", "53-58", "59-70").



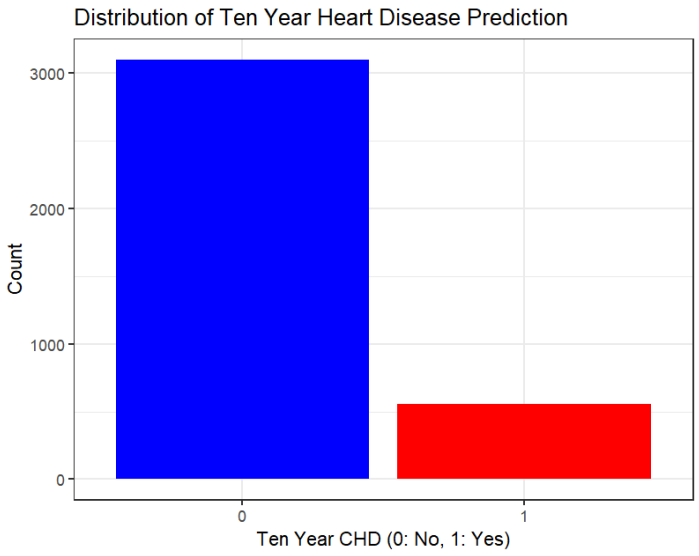
We made a model only using age as a predictor because we were interested to know how well you could predict someone's risk of heart disease by age alone and find a lethal dose for age.

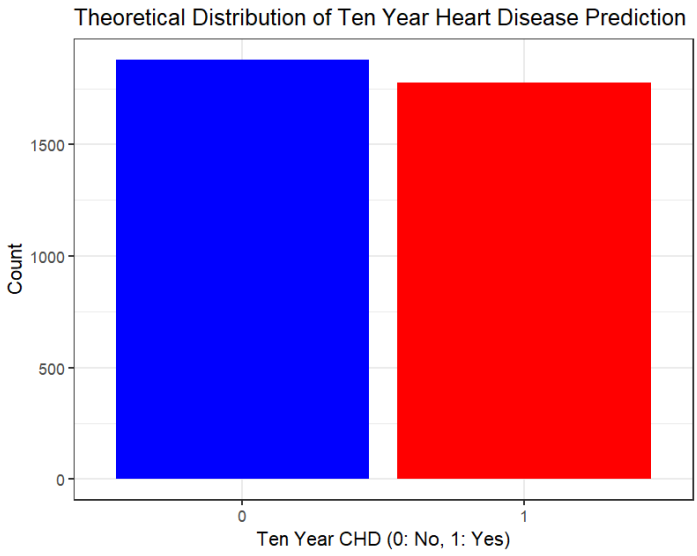
P-test	0 Predicted	1 Predicted
0 Actual	2882	217
1 Actual	458	458

$(2882 + 99)/(2882 + 217 + 458 + 99) = 0.81537$

$LD = 73.59727 \approx 74$

The accuracy of predicting heart disease with age alone as our explanatory variable was 81.5% accurate we did have to lower the value that the fitted values that are greater than 0.5 to 0.3 because the model wasn't predicting anyone to have heart disease but we are justified in lowering this value because we want to reduce the odds of a false positive that is we predict someone to not have heart disease when they actually do. The accuracy of the model with all of the variables in it was 0.8550328 Which is not much more than the model with just age alone and the model with age alone and we don't have a problem with multicollinearity with the age model. The lethal dose value of 74 means that at the age of 74, the odds that you have ten-year coronary heart disease is 50% or above.





The graphs above show the number of people who have coronary heart disease and don't have coronary heart disease. The graph on the left is the original data and as we can see there are not a lot of people who have coronary heart disease compared to the people who do not have coronary heart disease. The unbalanced data set may become biased toward the majority class, leading to poor performance on the minority class. To fix this we will use the ROSE function which stands for "Random Over-Sampling Examples", which is a technique used to address class imbalance issues in binary classification problems by generating synthetic data points to augment the minority class, essentially creating a more balanced dataset for machine learning models. The function utilizes a bootstrap-based approach to generate new data points by sampling from the minority class and creating new examples within its "neighborhood".

“ROSE tries to estimate the probability distribution P(x|y=k) for each class k and then draws the needed N_k samples from it. It’s well known that one way to estimate such density is through kernel density estimation which you can derive or intuit starting from more crude versions such as histogram analysis. The following describes KDE:

Given: data points x

Wanted: An estimate of P(x)

Operation: Choose a kernel function K(x) and then estimate P(x) as” (Wisam)

$$P(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(x - x_i\right)$$

In simpler words “ROSE samples the N_k points from this distribution once estimated for any class k (resulting in P(x|y=k)) and performing the following:

Choose a point randomly

Place the Gaussian on it

Sample one point from the Gaussian” (Wisam)

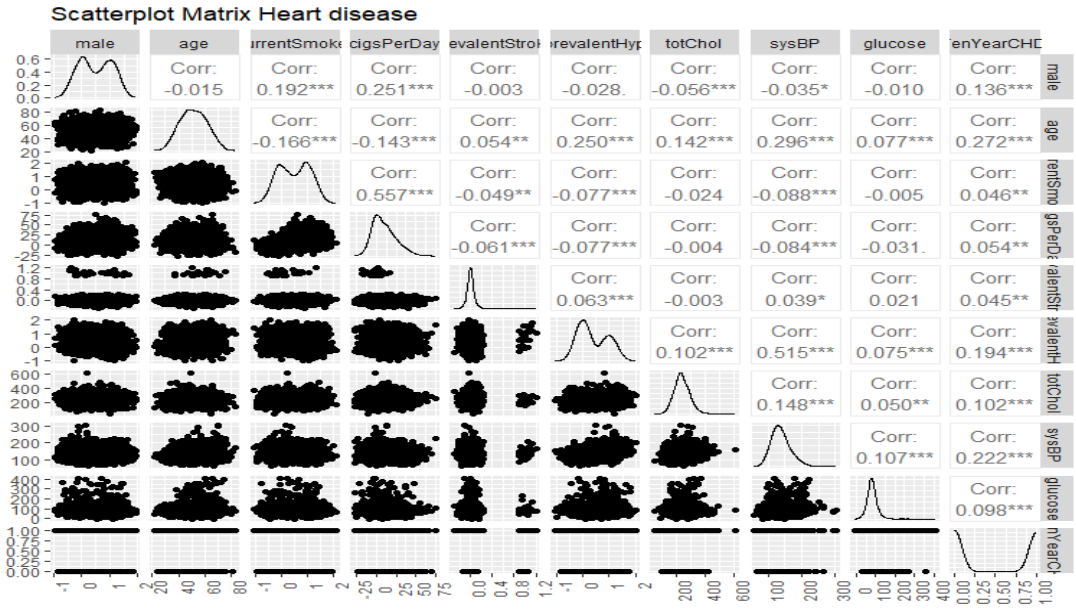
The graph on the left from before was the artificial data created by ROSE; we can see that it is much more balanced data. Now we will perform the same statistics that we did with our unbalanced data from before. We used the step function and again the backward and both directions agreed with each other the forward model had more variables so we went with the simpler model which was.

$$TenYearCHD = -5.3042137 + 0.4963144male + 0.0503652age + 0.2102881currentSmoker + 0.0066887cigsPerDay + 0.5575601prevalentStroke + 0.3050165prevalentHyp + 0.0022195totChol + 0.0090834sysBP + 0.0040708glucose$$

We also did the variance inflation factor for the balance data model and the VIF scores were much lower now and multicollinearity may be a problem But most of them were below five cigarettes per day and systolic blood pressure was about 7 and 6 respectively.

The slope interpretations For age: as age is increased by 1 the predicted odds that you have ten-year CHD is increased by $e^{0.0503652} = 1.051655$. For prevalentHyp: if you have prevalentHyp the predicted odd that you have ten-year CHD is increased by $e^{0.3050165} = 1.356647$.

The correlation matrix



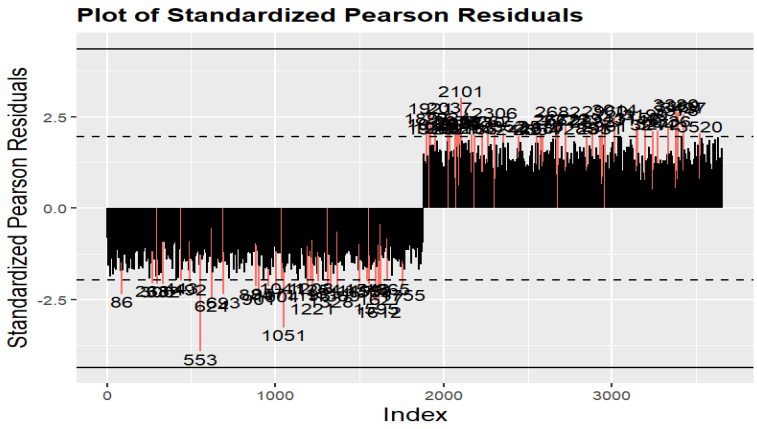
As we can see from the correlation Matrix systolic blood pressure and prevention hypertension are highly correlated with each other and Cigarettes per day and current smokers are highly correlated which were also the ones with the high VIF scores so that could be causing some of the multicollinearity. We could consider dropping one or two of these variables to reduce multicollinearity even more. The next thing we did was an ANOVA test.

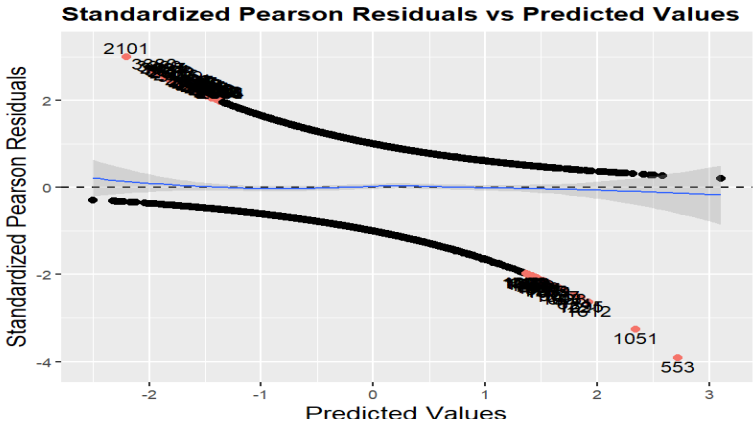
```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TenYearCHD

Terms added sequentially (first to last)

                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)

NULL                             3655     5065.6

male             1   68.455      3654     4997.1 < 2.2e-16 ***

age              1  289.152      3653     4707.9 < 2.2e-16 ***

currentSmoker    1   18.185      3652     4689.8 2.004e-05 ***
```

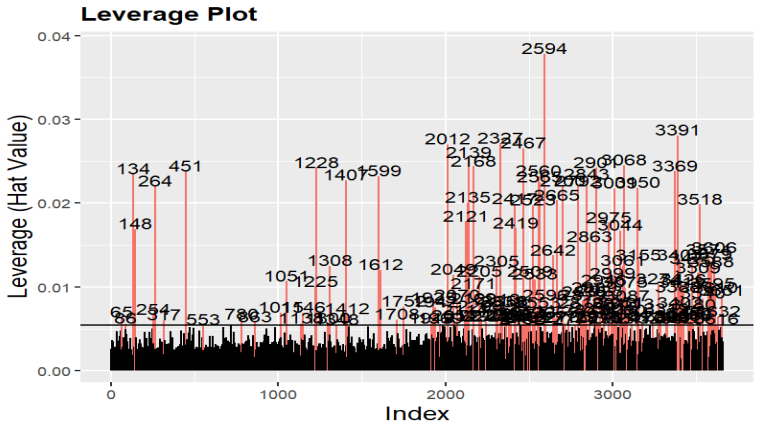
```
cigsPerDay        1    3.532       3651     4686.2   0.06019 .
prevalentStroke   1    4.700       3650     4681.5   0.03016 *
prevalentHyp      1   73.088       3649     4608.4 < 2.2e-16 ***
totChol           1   15.439       3648     4593.0 8.520e-05 ***
sysBP             1   39.736       3647     4553.3 2.906e-10 ***
glucose           1   16.451       3646     4536.8 4.992e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

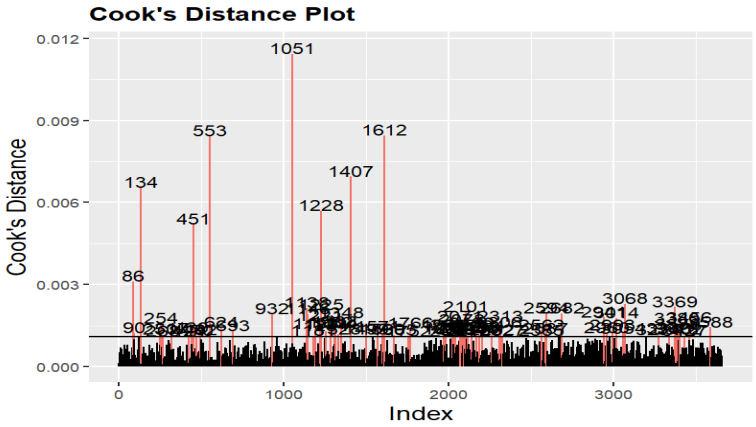
As we can see most of the variables significantly reduce the deviance in the model but it looks like cigarettes per day and prevalent stroke are not very significant this could give us another reason to remove these two variables which would help us reduce the multicollinearity.

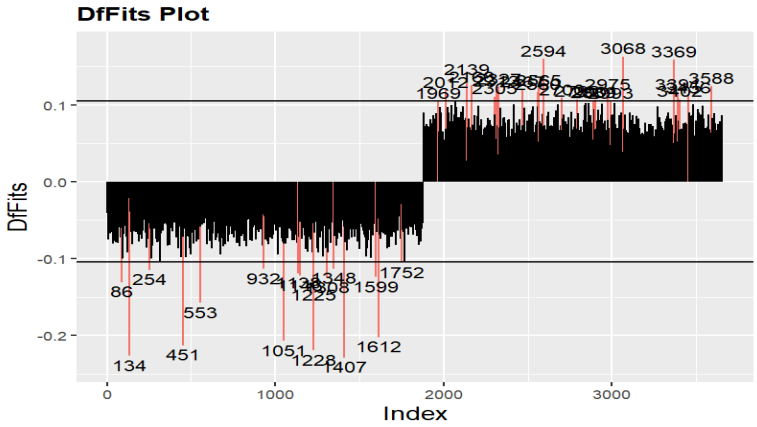
The next thing we did was model diagnostics with logistic plots.

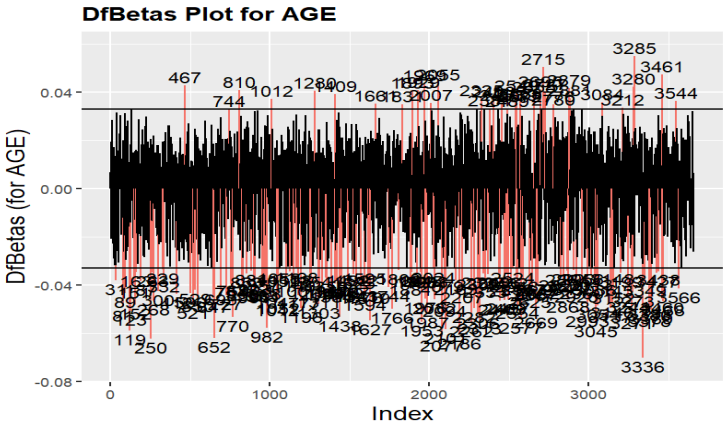




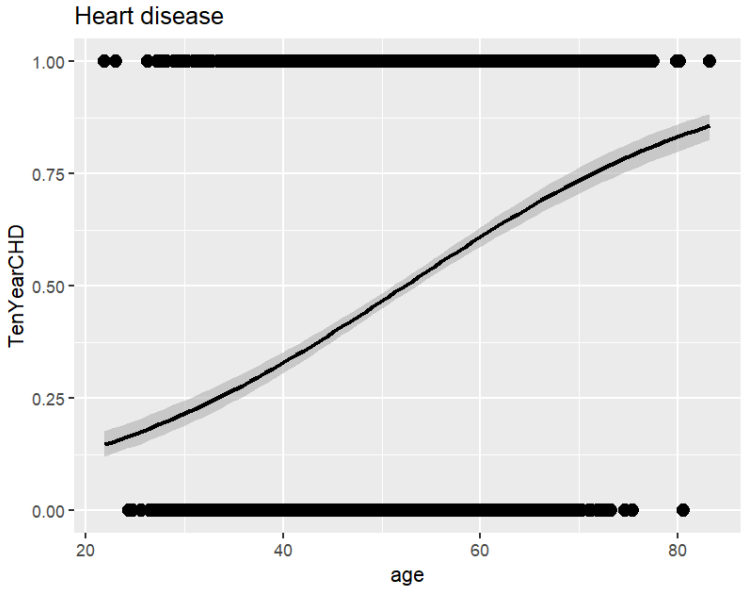


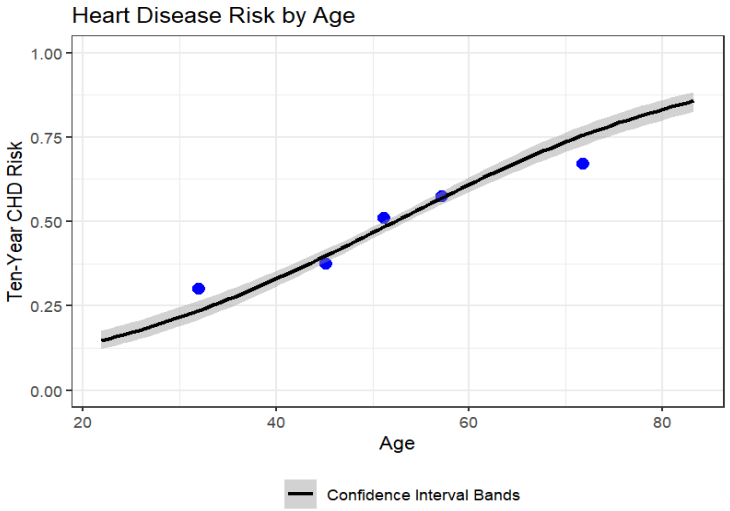






As we can see from the plots nothing seems too out of the ordinary and most importantly in the residual plot the smoothed conditional means line the blue one is mostly horizontal and at zero also the 95 confidence interval always captures zero in it as well which suggests that our model assumptions are met that is the variables are independent of one another and the relationship between them is linear.





Above are the logistic model plots for the balance data only using age as the explanatory variable. The one on the right is broken up into age groups ("22-42", "43-48", "49-54", "55-60", "61-83").

Accuracy of the model with the balanced data.

P-test	0 Predicted	1 Predicted
0 Actual	1241	637
1 Actual	742	1036

(1241 + 1036)/(1241 + 1036 + 742 + 637) = 0.6228118

LD = 52.29448

This is the accuracy of the model with only age as the explanatory variable as we can see the accuracy is 62.3%, and the accuracy of the model with all of the variables in it was 0.6619256 which isn't much more than the model with only age in it this could give us a reason to only use age as an explanatory variable because this would reduce multicollinearity it is a simpler model and still has accurate predictions. Also note that with this P-table, the fitted values were set greater than 0.5 whereas before it was 0.3 we didn't reduce it because we didn't need to like before but we still would be justified in reducing it to prevent a false positive. The lethal dose value for the age model was 52 this means at the age of 52 the predicted odds that you have coronary heart disease is greater than or equal to 50%.

Predictions, I predicted the probability that I have 10-year coronary heart disease using both models and my data. Using the model with the unbalanced data and all the variables

predicted odds that I had 10-year coronary heart disease was 0.01313699 and for the model with only age in it the predicted odds was 0.01532502. For the balance data set that had all the variables in it the predicted odds that I had 10-year coronary heart disease was 0.1091531 For the model with only age in it the predicted odds were 0.1339225. With both the models, we can see that I have a very low probability of having 10-year coronary heart disease but with the unbalanced data, the youngest person in the data set was 32 so predicting a 20-year-old (myself) we are extrapolating a little bit. With the balance data, the youngest person was 22 years old so we are no longer extrapolating but this is artificial data generated by the ROSE function.

If you would like to try to predict your odds of having 10-year coronary heart disease you can use the two models that were listed before if you don't know all of your data that is needed for the full models try using the models that only have age in them.

For original data: $TenYearCHD = -5.716299 + 0.077674 \cdot (age)$

For balanced data: $TenYearCHD = -3.022778 + 0.057803 \cdot (age)$

Just plug this into a calculator with your age in the age slot to find your predicted odds of having 10-year coronary heart disease.

In conclusion, significant predictors in demographics (age and gender) were highly significant predictors. Men appear to be more vulnerable to heart disease compared to women. Additionally, old age, higher daily cigarette consumption, and elevated systolic blood pressure are associated with increased odds of developing heart disease. Elevated glucose levels also contributed to the risk. Pre-existing conditions like hypertension was a notable risk factor. Our model seems to predict pretty good and the balance data also seems to predict well. Our model isn't perfect though we had to make artificial data to balance the data which may not represent

the population accurately. Multicollinearity could be a problem but the balance data doesn't have as big of a problem. Since the model is predicting Heart disease too many type II errors are not advisable aka a false negative predicting that someone doesn't have heart disease when they do is more dangerous than a false positive in this case. This means that we are justified in lowering the sensitivity threshold. Having more data in the category of people who have heart disease would be better so that we don't have to make artificial data. Overall the model is pretty good but could be improved with more data.

Works Cited

"Logistic regression To predict heart disease." *Kaggle*, https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data. Accessed 7 December 2024.

Wisam, Essam. "Class Imbalance: ROSE and Random Walk Oversampling (RWO)." *Towards Data Science*, 28 August 2023, https://towardsdatascience.com/class-imbalance-from-random-oversampling-to-rose-517e06d7a9b. Accessed 7 December 2024.

```
install.packages("remotes")
library(remotes)
install_github("rbrown53/math3150package")
library(math3150package)
install.packages("GGally")
library(GGally)
library(MASS)
install.packages("olsrr")
library(olsrr)
library(leaps)
library(MPV)
library(car)
library(caret)
library(dplyr)
install.packages("ROSE")
library(ROSE)
library(pROC)
library(faraway)
install.packages("nnet")
library(nnet)
#
================================================================================
=======

my_data <- data.frame(read.csv("C:/Users/david/Downloads/framingham.csv"))
head(my_data)

colSums(is.na(my_data))
my_data <- na.omit(my_data)
length(my_data$glucose)

my_mod <- glm(TenYearCHD ~ .,
              family = binomial, data = my_data)

step(my_mod, direction = "both")

mod_1 <- glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
               prevalentHyp + totChol + sysBP + glucose, family = binomial,
             data = my_data)

step(my_mod, direction = "backward")

# same as both or mod_1

step(my_mod, direction = "forward")
```

```
#mod_3 <-  glm(formula = TenYearCHD ~ male + age + education + currentSmoker +
#            cigsPerDay + BPMeds + prevalentStroke + prevalentHyp, +
#             diabetes + totChol + sysBP + diaBP + BMI + heartRate + glucose
#              family = binomial, data = my_data)+ diabetes +

summary(mod_1); vif(mod_1)


anova(mod_1, test = "Chisq")

ggpairs(data = my_data[, -c(3, 4, 6, 9, 12, 13, 14)],
        title = "Scatterplot Matrix Heart disease") +
  theme(axis.text.x = element_text(angle = 90))

logistic_plots(mod_1)


mod_1pred <- predict(mod_1, data.frame(male = 1, age = 20, cigsPerDay = 0,
             prevalentStroke = 0, prevalentHyp = 0, totChol = 145, sysBP = 116,
             glucose = 80))
ilogit(mod_1pred)

mod_age <- glm(TenYearCHD ~ age, data = my_data , family = binomial)
summary(mod_age)
ggplot(data = my_data, aes(y = TenYearCHD, x = age)) +
  geom_point(data = my_data, aes(y = TenYearCHD, x = age), size = 3) +
  labs(title = "Heart disease", y = "TenYearCHD", x = "age") +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
             formula = "y ~ x", se = T, color = "black",
             aes(fill = "Confidence Interval Bands"), data = my_data) +
  lims(y = c(0, 1)) +
  theme_bw() +
  scale_fill_manual("", values = "gray60") +
  theme(legend.position = "bottom")


breaks <- quantile(my_data$age, seq(0, 1, 0.2))  # Create quintiles for age
age_props <- my_data |>
  mutate(group = factor(
    case_when(
      age >= 32 & age <= 41 ~ "32-41",
      age >= 42 & age <= 46 ~ "42-46",
      age >= 47 & age <= 52 ~ "47-52",
      age >= 53 & age <= 58 ~ "53-58",
```

```
    age >= 59 & age <= 70 ~ "59-70",
    TRUE ~ NA_character_  # Catch-all for unexpected values
   ), levels = c("32-41", "42-46", "47-52", "53-58", "59-70")
  )) |>
  group_by(group) |>
  summarize(
   num_withCHD = sum(TenYearCHD == 1, na.rm = TRUE),
   num_withoutCHD = sum(TenYearCHD == 0, na.rm = TRUE),
   props = mean(TenYearCHD == 1, na.rm = TRUE)  # Proportion with CHD
  ) |>
  mutate(midpoints = breaks[1:5] + diff(breaks)/2)

print(age_props)

mod_age <- glm(TenYearCHD ~ age, data = my_data , family = binomial)
summary(mod_age)

ggplot(data = my_data, aes(y = TenYearCHD, x = age)) +
  geom_point(data = age_props, aes(y = props, x = midpoints),
          size = 3, color = "blue") +
  labs(title = "Heart Disease Risk by Age", y = "Ten-Year CHD Risk", x = "Age") +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
           formula = y ~ x, se = TRUE, color = "black",
           aes(fill = "Confidence Interval Bands")) +
  lims(y = c(0, 1)) +
  theme_bw() +
  scale_fill_manual("", values = "gray60") +
  theme(legend.position = "bottom")

LD <- (5.71630/0.07767)
LD

ptest <- as.numeric(mod_age$fitted.values > 0.3)

table(my_data$TenYearCHD, ptest)

ptest_1 <- as.numeric(mod_1$fitted.values > 0.5)

table(my_data$TenYearCHD, ptest_1)


ggplot(my_data, aes(x = factor(TenYearCHD))) +
  geom_bar(fill = c("blue", "red")) +
  labs(
    title = "Distribution of Ten Year Heart Disease Prediction",
```

```
    x = "Ten Year CHD (0: No, 1: Yes)",
    y = "Count"
  ) +
  theme_bw()

mod_age_pred <- predict(mod_age, data.frame(age = 20))
ilogit(mod_age_pred)

#=============================================================
===============

balanced_data <- ROSE(TenYearCHD ~ ., data = my_data, seed = 1)$data
head(balanced_data)
New_age_mod <- glm(TenYearCHD ~ age, data = balanced_data, family = binomial)
summary(New_age_mod)



ggplot(data = balanced_data, aes(x = factor(TenYearCHD))) +
  geom_bar(fill = c("blue", "red")) +
  labs(
    title = "Theoretical Distribution of Ten Year Heart Disease Prediction",
    x = "Ten Year CHD (0: No, 1: Yes)",
    y = "Count"
  ) +
  theme_bw()


balanced_mod <- glm(TenYearCHD ~ ., family = binomial, data = balanced_data)

step(balanced_mod, direction = "both")
#balanced_mod_1 <- glm(formula = TenYearCHD ~ male + age + currentSmoker +
#                cigsPerDay + prevalentStroke + prevalentHyp + totChol +
#                 sysBP + glucose, family = binomial, data = balanced_data)

step(balanced_mod, direction = "forward") # to many don't use


step(balanced_mod, direction = "backward") # same as both


balanced_mod_filtered <- glm(formula = TenYearCHD ~ male + age + currentSmoker +
                       cigsPerDay + prevalentStroke + prevalentHyp + totChol
                       + sysBP + glucose,
                       family = binomial, data = balanced_data)
```

```
summary(balanced_mod_filtered); vif(balanced_mod_filtered)

balanced_mod_filtered_pred <- predict(balanced_mod_filtered,
                                      data.frame(male = 1, age = 20,
  cigsPerDay = 0, currentSmoker = 0, prevalentStroke = 0, prevalentHyp = 0,
  totChol = 145, sysBP = 116, glucose = 80))

ilogit(balanced_mod_filtered_pred)

ptest <- as.numeric(balanced_mod_filtered$fitted.values > 0.5)

table(balanced_data$TenYearCHD, ptest)

(1299 + 1121)/(1299 + 579 + 657 + 1121)

anova(balanced_mod_filtered, test = "Chisq")

ggpairs(data = balanced_data[, -c(3, 6, 9, 12, 13, 14)],
        title = "Scatterplot Matrix Heart disease") +
  theme(axis.text.x = element_text(angle = 90))

logistic_plots(balanced_mod_filtered)


balanced_mod_age <- glm(TenYearCHD ~ age, data = balanced_data , family =
binomial)
summary(balanced_mod_age); vif(balanced_mod_age)


ggplot(data = balanced_data, aes(y = TenYearCHD, x = age)) +
  geom_point(data = balanced_data, aes(y = TenYearCHD, x = age), size = 3) +
  labs(title = "Heart disease", y = "TenYearCHD", x = "age") +
  geom_smooth(
    method = "glm",
    method.args = list(family = "binomial"),
    formula = y ~ x,
    se = TRUE,
    color = "black",
    fill = "gray60",  # Replace aesthetic mapping with a fixed value
    data = balanced_data
  )

breaks <- quantile(balanced_data$age, seq(0, 1, 0.2))  # Create quintiles for age
age_props_balanced <- balanced_data |>
```

```
  mutate(group = factor(
    case_when(
      age >= breaks[1] & age < breaks[2] ~ "22-42",
      age >= breaks[2] & age < breaks[3] ~ "43-48",
      age >= breaks[3] & age < breaks[4] ~ "49-54",
      age >= breaks[4] & age < breaks[5] ~ "55-60",
      age >= breaks[5] & age <= breaks[6] ~ "61-83",
      TRUE ~ NA_character_  # Catch-all for unexpected values
    ), levels = c("22-42", "43-48", "49-54", "55-60", "61-83")
  )) |>
  group_by(group) |>
  summarize(
    num_withCHD = sum(TenYearCHD == 1, na.rm = TRUE),
    num_withoutCHD = sum(TenYearCHD == 0, na.rm = TRUE),
    props = mean(TenYearCHD == 1, na.rm = TRUE)  # Proportion with CHD
  ) |>
  mutate(midpoints = (breaks[1:5] + breaks[2:6]) / 2)

print(age_props_balanced)

ggplot(data = balanced_data, aes(y = TenYearCHD, x = age)) +
  geom_point(data = age_props_balanced, aes(y = props, x = midpoints),
             size = 3, color = "blue") +
  labs(title = "Heart Disease Risk by Age", y = "Ten-Year CHD Risk", x = "Age") +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
              formula = y ~ x, se = TRUE, color = "black",
              aes(fill = "Confidence Interval Bands")) +
  lims(y = c(0, 1)) +
  theme_bw() +
  scale_fill_manual("", values = "gray60") +
  theme(legend.position = "bottom")


balanced_mod_age_pred <- predict(balanced_mod_age,
                                 data.frame(age = 20))
ilogit(balanced_mod_age_pred)

LD_2 <- (3.022778/0.057803)
LD_2

ptest <- as.numeric(New_age_mod$fitted.values > 0.5)

table(balanced_data$TenYearCHD, ptest)

(1241 + 1036)/(1241 + 1036 + 742 + 637)
```