Regression II Final Project Group 1M

Yingzhi Ma, Guanyu Lu, Yi Yang

2023-11-26

Package may use in the project

```
library(pander)
library(ggplot2)
library(moments)
library(tidyverse)
library(psych)
library(rio)
library(MASS)
library(ResourceSelection)
library(car)
library(VGAM)
library(pROC)
library(lmtest)
```

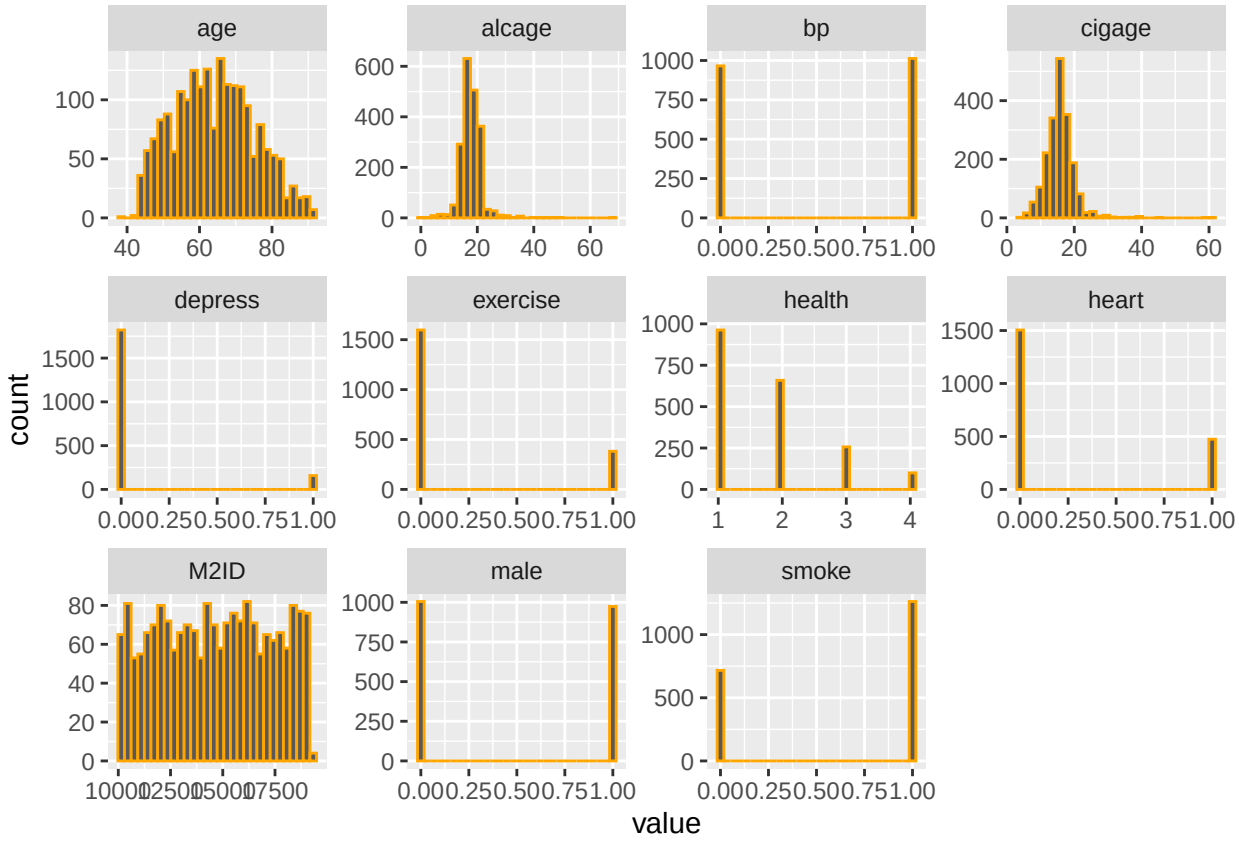
Upload Data from Excel File

```
my_data <- read.csv("MIDUS_III_Final_Exam_Fall2023_data.csv")
#Show the first six rows of data set.
pander(head(my_data))
```

M2ID	age	male	heart	cigage	smoke	alcage	depress	bp	exercise	health
10001	69	1	0	13	1	18	0	1	0	2
10015	63	0	1	15	1	20	1	1	1	3
10024	60	1	0	12	0	18	1	0	0	2
10037	51	1	1	12	1	13	0	0	1	3
10038	66	0	1	10	0	22	0	0	1	2
10040	58	1	0	13	1	13	0	0	0	1

```
my_data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(col = 'orange')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Show the distribution of data set.
pander(skewness(my_data), caption = 'Skewness of numeric data')
```

Table 2: Table continues below

M2ID	age	male	heart	cigage	smoke	alcage	depress
-0.01671	0.1646	0.03133	1.221	2.02	-0.5729	2.529	3.1

bp	exercise	health
-0.04751	1.556	0.9821

```
# Check Correlation
cor <- cor(my_data);
pander(cor);
```

Table 4: Table continues below

	M2ID	age	male	heart	cigage
M2ID	1	-0.01956	-0.02841	-0.01104	0.02814
age	-0.01956	1	0.06159	0.2291	0.1652
male	-0.02841	0.06159	1	0.1035	-0.1755

	M2ID	age	male	heart	cigage
heart	-0.01104	0.2291	0.1035	1	-0.05544
cigage	0.02814	0.1652	-0.1755	-0.05544	1
smoke	-0.01441	0.07454	0.008168	0.07816	0.02053
alcage	0.01446	0.2318	-0.2018	-0.001475	0.3084
depress	0.01847	-0.09125	-0.1222	0.02253	-0.02107
bp	-0.01194	0.2289	0.0393	0.2117	0.0341
exercise	-0.01475	-0.06617	-0.1076	-0.007484	-0.03271
health	-0.003769	0.06318	-0.03024	0.2335	0.01789

	smoke	alcage	depress	bp	exercise	health
M2ID	-0.01441	0.01446	0.01847	-0.01194	-0.01475	-0.003769
age	0.07454	0.2318	-0.09125	0.2289	-0.06617	0.06318
male	0.008168	-0.2018	-0.1222	0.0393	-0.1076	-0.03024
heart	0.07816	-0.001475	0.02253	0.2117	-0.007484	0.2335
cigage	0.02053	0.3084	-0.02107	0.0341	-0.03271	0.01789
smoke	1	-0.03117	0.02422	0.09677	-0.04954	0.1268
alcage	-0.03117	1	0.001663	0.08179	-0.05251	0.021
depress	0.02422	0.001663	1	0.004191	0.05433	0.1749
bp	0.09677	0.08179	0.004191	1	-0.07821	0.2458
exercise	-0.04954	-0.05251	0.05433	-0.07821	1	-0.04917
health	0.1268	0.021	0.1749	0.2458	-0.04917	1

Univariate Analysis

```
# Continuous Variables
cont_var <- my_data %>% dplyr::select(age, cigage, alcage)
summary(cont_var)
```

```
##       age             cigage          alcage
##  Min.   :39.00   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:55.00   1st Qu.:13.00   1st Qu.:16.00
##  Median :64.00   Median :16.00   Median :17.00
##  Mean   :64.09   Mean   :15.63   Mean   :17.66
##  3rd Qu.:72.00   3rd Qu.:18.00   3rd Qu.:19.00
##  Max.   :92.00   Max.   :60.00   Max.   :69.00
```

```
# Categorical Variables
# List of categorical variables
cat.var.Final <- c("male", "heart", "smoke", "depress", "bp", "exercise", "health")

# Loop through each categorical variable and print its percentage frequency table
for (var in cat.var.Final) {
  cat("\nPercentage Frequency table for", var, ":\n")
  freq.table <- table(my_data[[var]])
  perc.table <- round((freq.table / sum(freq.table)) * 100, 2) # calculate percentages and round off to
  pander(freq.table)
  pander(perc.table)
}
```

```
##
## Percentage Frequency table for male :
##
## Percentage Frequency table for heart :
##
## Percentage Frequency table for smoke :
##
## Percentage Frequency table for depress :
##
## Percentage Frequency table for bp :
##
## Percentage Frequency table for exercise :
##
## Percentage Frequency table for health :
```

Bivariate Analysis

```
# Case 1
## T - Test for Continuous Predictors
t.test(my_data$heart,my_data$alcage)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$heart and my_data$alcage
## t = -201.36, df = 2027.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.59212 -17.25275
## sample estimates:
##  mean of x  mean of y
##  0.2395149 17.6619505
```

```
t.test(my_data$heart,my_data$cigage)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$heart and my_data$cigage
## t = -157.27, df = 2016.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.58508 -15.20117
## sample estimates:
##  mean of x  mean of y
##  0.2395149 15.6326427
```

```
t.test(my_data$heart,my_data$age)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$heart and my_data$age
## t = -257.73, df = 1983.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -64.34088 -63.36908
## sample estimates:
##  mean of x  mean of y
##  0.2395149 64.0944922
```

```r
# Calculating mean and SD for each continuous predictor within
# each category of 'heart' for 'age'
age_stat <- my_data %>%
  group_by(heart) %>%
  summarise(
    Mean_Age = mean(age, na.rm = TRUE),
    SD_Age = sd(age, na.rm = TRUE)
  )
print(age_stat)
```

```
## # A tibble: 2 x 3
##   heart Mean_Age SD_Age
##   <int>    <dbl>  <dbl>
## 1     0     62.7   10.7
## 2     1     68.6   10.7
```

```r
# for 'cigage'
cigage_stat <- my_data %>%
  group_by(heart) %>%
  summarise(
    Mean_Cigage = mean(cigage, na.rm = TRUE),
    SD_Cigage = sd(cigage, na.rm = TRUE)
  )
print(cigage_stat)
```

```
## # A tibble: 2 x 3
##   heart Mean_Cigage SD_Cigage
##   <int>       <dbl>     <dbl>
## 1     0        15.8      4.30
## 2     1        15.2      4.42
```

```r
# for 'alcage'
alcage_stat <- my_data %>%
  group_by(heart) %>%
  summarise(
    Mean_Alcage = mean(alcage, na.rm = TRUE),
    SD_Alcage = sd(alcage, na.rm = TRUE)
  )
print(alcage_stat)
```

```
## # A tibble: 2 x 3
##   heart Mean_Alcage SD_Alcage
##   <int>       <dbl>     <dbl>
## 1     0        17.7      3.65
## 2     1        17.7      4.34
```

```
## Chi-Squared Tests for Categorical Predictors
### Loop through each categorical predictor and create a contingency table with 'heart'
for (predictor in cat.var.Final) {
  cat("\nContingency table for", predictor, "and heart:\n")
  conti.table.1 <- table(my_data[[predictor]], my_data$heart)
  perc.conti.table.1 <- prop.table(conti.table.1, margin = 2) * 100
  chi.squared.test.1 <- chisq.test(conti.table.1)
  print(chi.squared.test.1)
  print(conti.table.1)
  print(perc.conti.table.1)
}
```

```
##
## Contingency table for male and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conti.table.1
## X-squared = 20.726, df = 1, p-value = 5.299e-06
##
##
##       0   1
##   0 808 197
##   1 697 277
##
##           0        1
##   0 53.68771 41.56118
##   1 46.31229 58.43882
##
## Contingency table for heart and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conti.table.1
## X-squared = 1973.5, df = 1, p-value < 2.2e-16
##
##
##       0    1
##   0 1505    0
##   1    0  474
##
##       0   1
##   0 100   0
##   1   0 100
##
## Contingency table for smoke and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  conti.table.1
## X-squared = 11.712, df = 1, p-value = 0.0006208
##
##
##       0   1
##   0 577 140
##   1 928 334
##
##           0        1
##   0 38.33887 29.53586
##   1 61.66113 70.46414
##
## Contingency table for depress and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conti.table.1
## X-squared = 0.81885, df = 1, p-value = 0.3655
##
##
##        0    1
##   0 1390  431
##   1  115   43
##
##             0         1
##   0 92.358804 90.928270
##   1  7.641196  9.071730
##
## Contingency table for bp and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conti.table.1
## X-squared = 87.692, df = 1, p-value < 2.2e-16
##
##
##       0   1
##   0 824 142
##   1 681 332
##
##           0        1
##   0 54.75083 29.95781
##   1 45.24917 70.04219
##
## Contingency table for exercise and heart:
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conti.table.1
## X-squared = 0.070861, df = 1, p-value = 0.7901
##
##
##        0    1
```

```
##   0 1212  385
##   1  293   89
##
##           0        1
##   0 80.53156 81.22363
##   1 19.46844 18.77637
##
## Contingency table for health and heart:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.1
## X-squared = 108.69, df = 3, p-value < 2.2e-16
##
##
##       0   1
##   1 817 146
##   2 480 179
##   3 155 102
##   4  53  47
##
##             0         1
##   1 54.285714 30.801688
##   2 31.893688 37.763713
##   3 10.299003 21.518987
##   4  3.521595  9.915612
```

```r
# Case 2
## Logistic Regression for Continuous Predictors
# T - Test for Continuous Predictors with 'health' as Outcome
t.test(my_data$health,my_data$alcage)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$health and my_data$alcage
## t = -180.52, df = 2181.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.09055 -15.74472
## sample estimates:
## mean of x mean of y
##  1.744315 17.661950
```

```r
t.test(my_data$health,my_data$cigage)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$health and my_data$cigage
## t = -139.8, df = 2136.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -14.08315 -13.69351
## sample estimates:
## mean of x mean of y
##  1.744315 15.632643
```

```
t.test(my_data$health,my_data$age)
```

```
##
##  Welch Two Sample t-test
##
## data:  my_data$health and my_data$age
## t = -251.06, df = 2002.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -62.83722 -61.86314
## sample estimates:
## mean of x mean of y
##  1.744315 64.094492
```

```
# Calculating mean and SD for each continuous predictor within each category of 'health'
# for 'age'
age_stat_2 <- my_data %>%
  group_by(health) %>%
  summarise(
    Mean_Age = mean(age, na.rm = TRUE),
    SD_Age = sd(age, na.rm = TRUE)
  )
print("Age statistics by health categories:")
```

```
## [1] "Age statistics by health categories:"
```

```
print(age_stat_2)
```

```
## # A tibble: 4 x 3
##   health Mean_Age SD_Age
##    <int>    <dbl>  <dbl>
## 1      1     63.3   10.6
## 2      2     64.7   11.2
## 3      3     65.6   11.2
## 4      4     64.0   12.3
```

```
# for 'cigage'
cigage_stat_2 <- my_data %>%
  group_by(health) %>%
  summarise(
    Mean_Cigage = mean(cigage, na.rm = TRUE),
    SD_Cigage = sd(cigage, na.rm = TRUE)
  )
print("Cigage statistics by health categories:")
```

```
## [1] "Cigage statistics by health categories:"
```

```r
print(cigage_stat_2)
```

```
## # A tibble: 4 x 3
##   health Mean_Cigage SD_Cigage
##    <int>       <dbl>     <dbl>
## 1      1        15.6      3.96
## 2      2        15.7      4.33
## 3      3        15.6      4.72
## 4      4        15.9      6.34
```

```r
# for 'alcage'
alcage_stat_2 <- my_data %>%
  group_by(health) %>%
  summarise(
    Mean_Alcage = mean(alcage, na.rm = TRUE),
    SD_Alcage = sd(alcage, na.rm = TRUE)
  )
print("Alcage statistics by health categories:")
```

```
## [1] "Alcage statistics by health categories:"
```

```r
print(alcage_stat_2)
```

```
## # A tibble: 4 x 3
##   health Mean_Alcage SD_Alcage
##    <int>       <dbl>     <dbl>
## 1      1        17.6      3.26
## 2      2        17.5      3.27
## 3      3        18.1      5.26
## 4      4        17.7      6.77
```

```r
## Chi-Squared Tests for Categorical Predictors
# Loop through each categorical predictor and conduct a Chi-squared test with 'health'
for (predictor in cat.var.Final) {
  cat("\nContingency table for", predictor, "and health:\n")
  conti.table.2 <- table(my_data[[predictor]], my_data$health)
  perc.conti.table.2 <- prop.table(conti.table.2, margin = 2) * 100
  chi.squared.test.2 <- chisq.test(conti.table.2)
  print(chi.squared.test.2)
  print(conti.table.2)
  print(perc.conti.table.2)
}
```

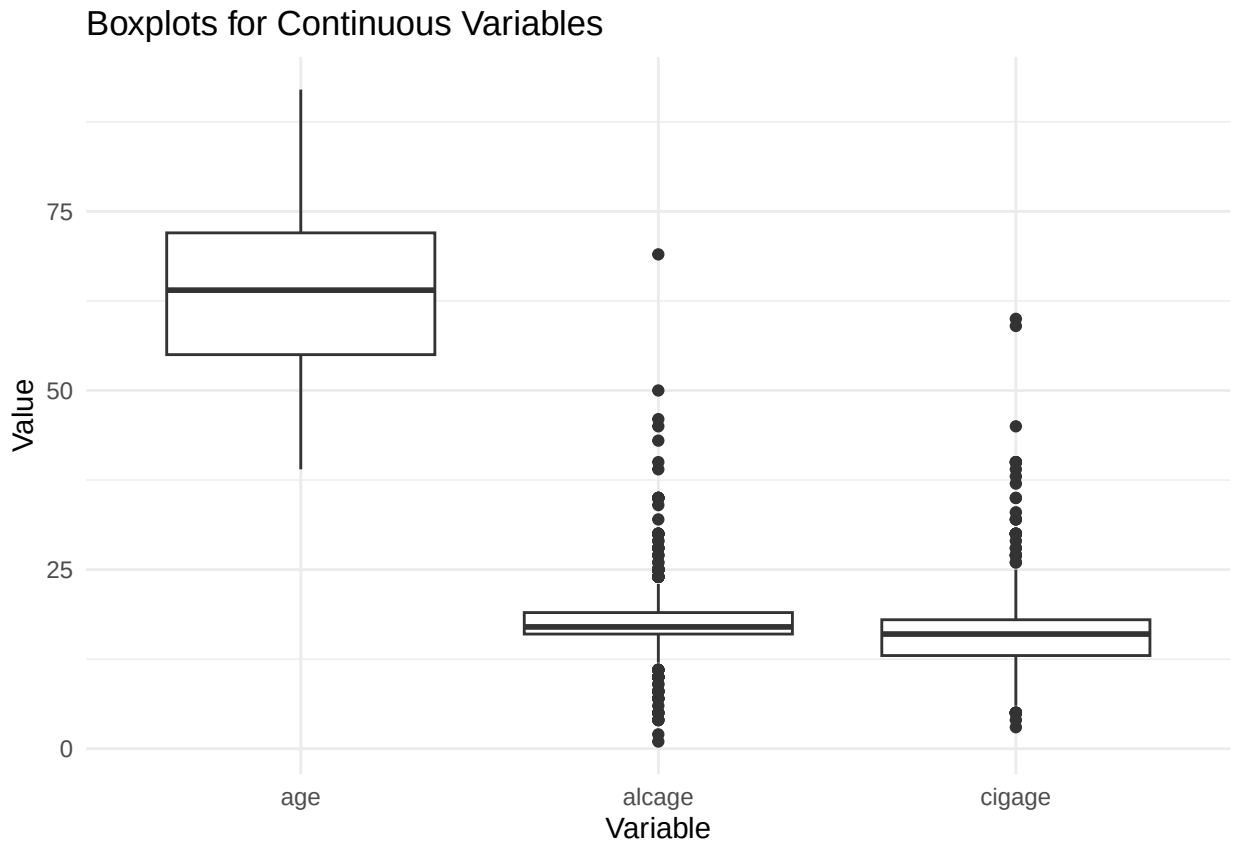
```
##
## Contingency table for male and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 2.3622, df = 3, p-value = 0.5007
##
```

```
##
##       1   2   3   4
##   0 481 330 138  56
##   1 482 329 119  44
##
##           1        2        3        4
##   0 49.94808 50.07587 53.69650 56.00000
##   1 50.05192 49.92413 46.30350 44.00000
##
## Contingency table for heart and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 108.69, df = 3, p-value < 2.2e-16
##
##
##       1   2   3   4
##   0 817 480 155  53
##   1 146 179 102  47
##
##           1        2        3        4
##   0 84.83904 72.83763 60.31128 53.00000
##   1 15.16096 27.16237 39.68872 47.00000
##
## Contingency table for smoke and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 37.437, df = 3, p-value = 3.72e-08
##
##
##       1   2   3   4
##   0 405 224  59  29
##   1 558 435 198  71
##
##           1        2        3        4
##   0 42.05607 33.99090 22.95720 29.00000
##   1 57.94393 66.00910 77.04280 71.00000
##
## Contingency table for depress and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 75.94, df = 3, p-value = 2.278e-16
##
##
##       1   2   3   4
##   0 913 619 212  77
##   1  50  40  45  23
##
##             1         2         3         4
```

```
##   0 94.807892 93.930197 82.490272 77.000000
##   1  5.192108  6.069803 17.509728 23.000000
##
## Contingency table for bp and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 122.24, df = 3, p-value < 2.2e-16
##
##
##       1   2   3   4
##   0 581 281  85  19
##   1 382 378 172  81
##
##           1        2        3        4
##   0 60.33229 42.64036 33.07393 19.00000
##   1 39.66771 57.35964 66.92607 81.00000
##
## Contingency table for exercise and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 4.851, df = 3, p-value = 0.183
##
##
##       1   2   3   4
##   0 760 537 215  85
##   1 203 122  42  15
##
##           1        2        3        4
##   0 78.92004 81.48710 83.65759 85.00000
##   1 21.07996 18.51290 16.34241 15.00000
##
## Contingency table for health and health:
##
##  Pearson's Chi-squared test
##
## data:  conti.table.2
## X-squared = 5937, df = 9, p-value < 2.2e-16
##
##
##       1   2   3   4
##   1 963   0   0   0
##   2   0 659   0   0
##   3   0   0 257   0
##   4   0   0   0 100
##
##       1   2   3   4
##   1 100   0   0   0
##   2   0 100   0   0
##   3   0   0 100   0
##   4   0   0   0 100
```

Assumptions

```
# Checking for outliers in continuous variables
cont_var_long <- pivot_longer(my_data, cols = c(age, cigage, alcage),
                                    names_to = "variable", values_to = "value")

## Create boxplots for age, cigage, and alcage
ggplot(cont_var_long, aes(x = variable, y = value)) +
    geom_boxplot() +
    theme_minimal() +
    ggtitle("Boxplots for Continuous Variables") +
    xlab("Variable") +
    ylab("Value")
```




```
# Checking for data consistency
summary(cont_var)
```

```
##       age            cigage          alcage
##  Min.   :39.00   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:55.00   1st Qu.:13.00   1st Qu.:16.00
##  Median :64.00   Median :16.00   Median :17.00
##  Mean   :64.09   Mean   :15.63   Mean   :17.66
##  3rd Qu.:72.00   3rd Qu.:18.00   3rd Qu.:19.00
##  Max.   :92.00   Max.   :60.00   Max.   :69.00
```

```r
# Checking for Multicollinearity
## Correlation matrix for 'age', 'cigage', and 'alcage'
cor <- cor(my_data[c("age", "cigage", "alcage")])
print(cor)
```

```
##              age    cigage    alcage
## age    1.0000000 0.1652277 0.2318361
## cigage 0.1652277 1.0000000 0.3084107
## alcage 0.2318361 0.3084107 1.0000000
```

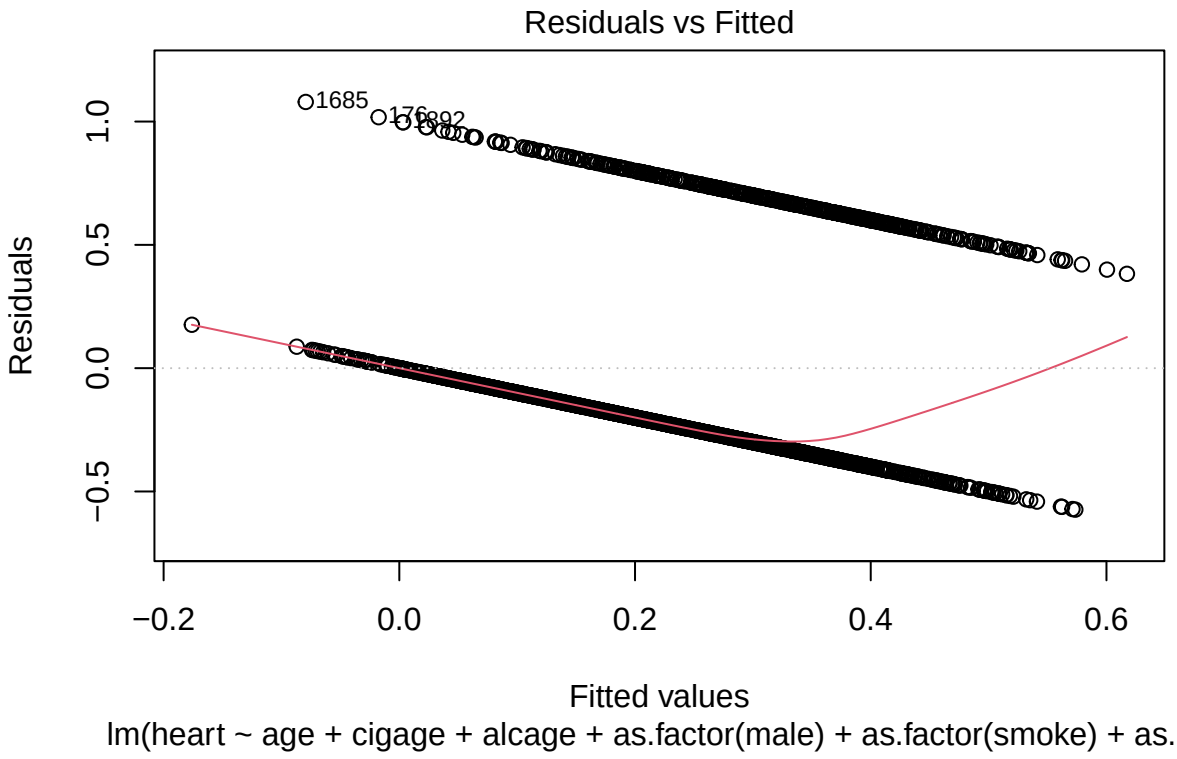
```r
my_data_2_3 <- my_data %>% filter(age == 64) %>% filter(bp == 1);
## Standardization for Continuous Variables in case there are multicollinearity
my_data$age <- scale(my_data$age)
my_data$cigage <- scale(my_data$cigage)
my_data$alcage <- scale(my_data$alcage)

## Case 1
### Use Linear model as a proxy
linear_model_assump <- lm(heart ~ age + cigage + alcage + as.factor(male) +
                            as.factor(smoke) + as.factor(depress) +
                            as.factor(bp) + as.factor(exercise), data = my_data)
pander(vif(linear_model_assump))
```

Table 6: Table continues below

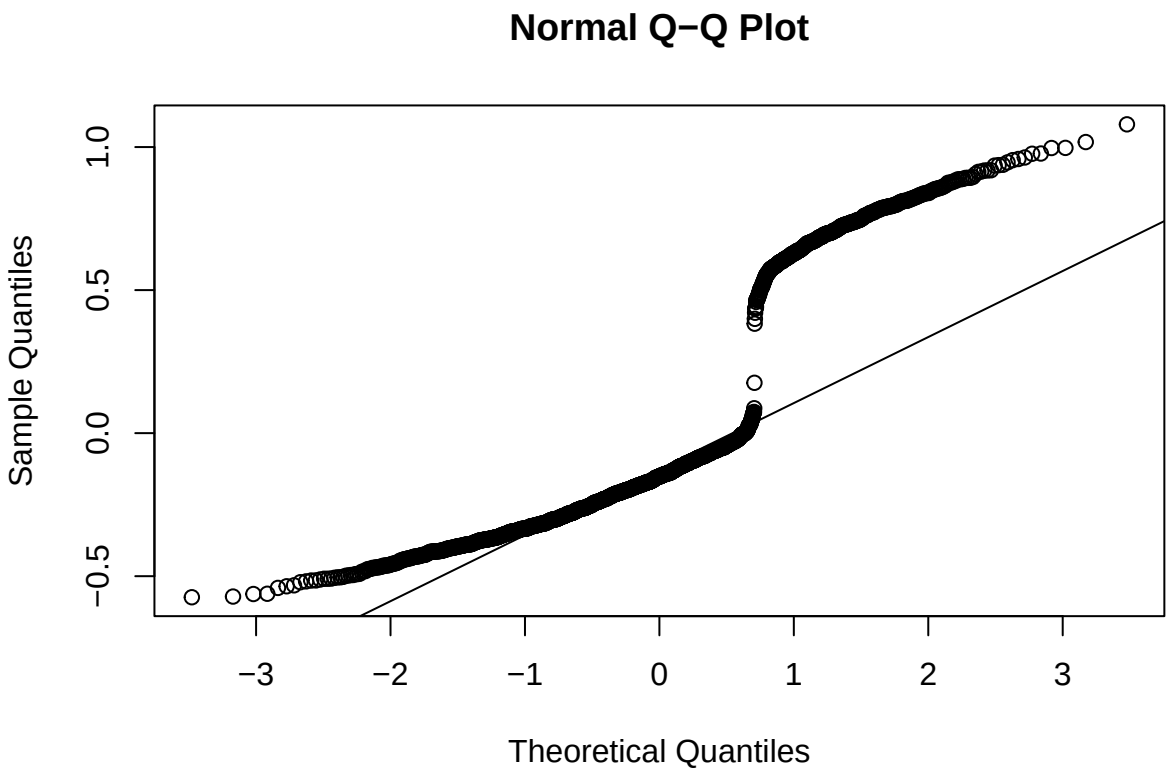
age	cigage	alcage	as.factor(male)	as.factor(smoke)
1.149	1.14	1.195	1.106	1.019

as.factor(depress)	as.factor(bp)	as.factor(exercise)
1.027	1.069	1.027

```r
## Homoscedasticity
plot(linear_model_assump, which = 1)
```




```
# Normality of Residuals
qqnorm(residuals(linear_model_assump))
qqline(residuals(linear_model_assump))
```




```
# Independence of Errors
dwtest(linear_model_assump)
```

```
##
##  Durbin-Watson test
##
## data:  linear_model_assump
## DW = 2.0712, p-value = 0.9432
## alternative hypothesis: true autocorrelation is greater than 0
```

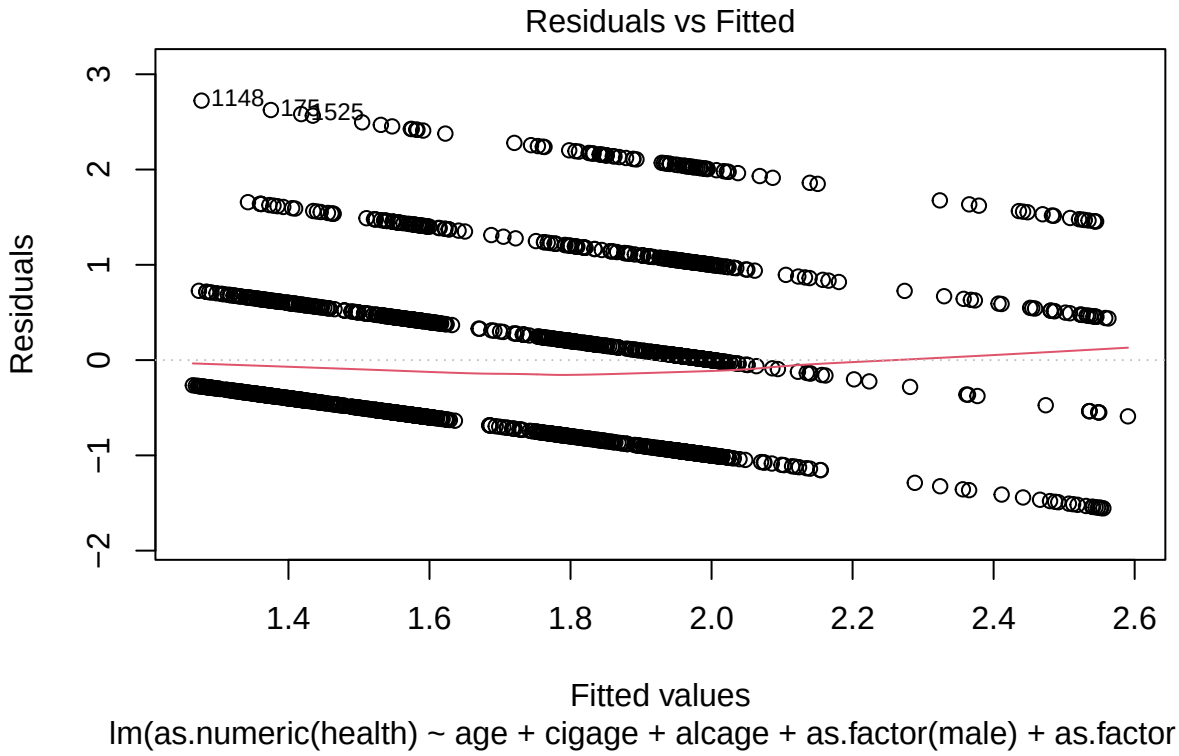
```
## Case 2
### Use Linear model as a proxy
linear_model_assump_2 <- lm(as.numeric(health) ~ age + cigage +
                              alcage + as.factor(male) + as.factor(smoke) +
                              as.factor(depress) + as.factor(bp) +
                              as.factor(exercise), data = my_data)
pander(vif(linear_model_assump_2))
```

Table 8: Table continues below

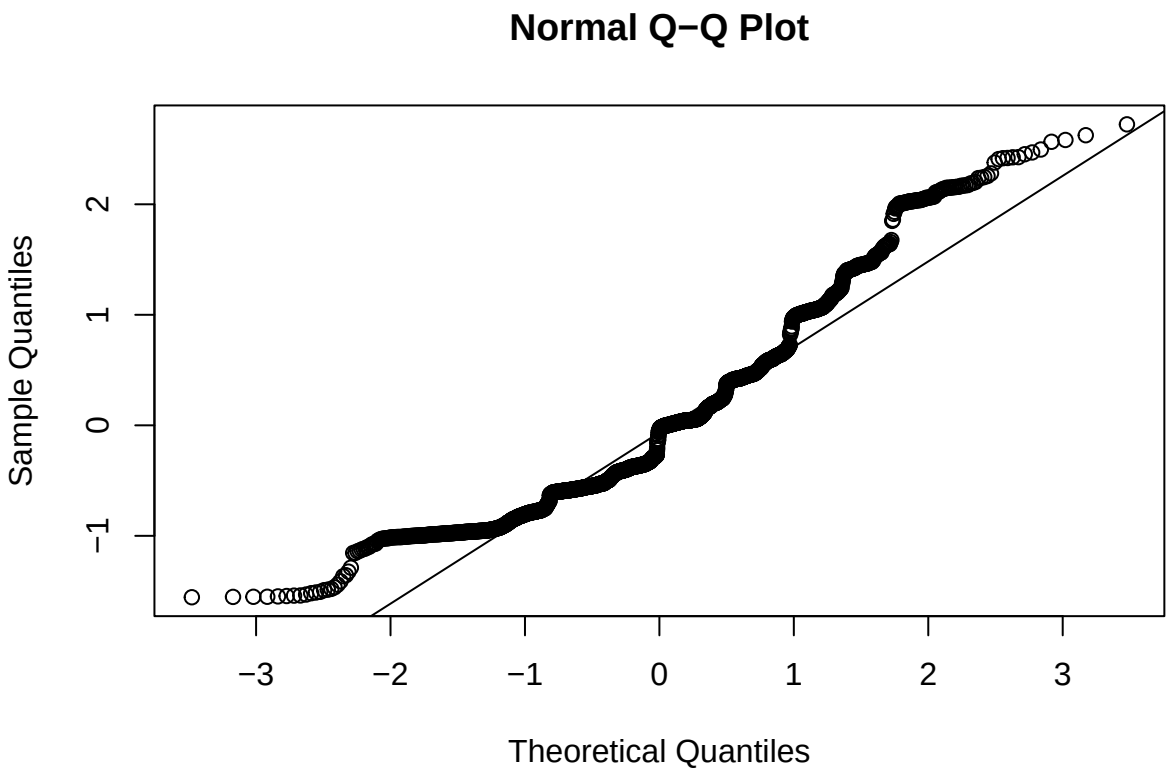
age	cigage	alcage	as.factor(male)	as.factor(smoke)
1.149	1.14	1.195	1.106	1.019

as.factor(depress)	as.factor(bp)	as.factor(exercise)
1.027	1.069	1.027

```
## Homoscedasticity
plot(linear_model_assump_2, which = 1)
```




```
# Normality of Residuals
qqnorm(residuals(linear_model_assump_2))
qqline(residuals(linear_model_assump_2))
```




```
# Independence of Errors
dwtest(linear_model_assump_2)
```

```
##
##  Durbin-Watson test
##
## data:  linear_model_assump_2
## DW = 2.0817, p-value = 0.9653
## alternative hypothesis: true autocorrelation is greater than 0
```

Case1

For the first study, you will build a predictive model for predicting if people in their midlife have ever experienced heart trouble (outcome variable). Step 1: select the correct model based on the distribution of the outcome variable. You might consider the following independent variables (bp, smoke, age, male, and exercise) as potential predictors. In step 2, you will run the model with the interaction term between smoke and male, controlling for the other variables. In step 3, you will assess if each model is a good fit for the data and which model (the main effects or interaction effect model) is better. Remember data cleaning and checking for potential outliers that might influence the estimates in the model is one of the major steps in statistical analysis. (NOTE: DO NOT DELETE ANY OUTLIERS, NOTE AND EVALUATE THEM).

```
# Step 1
my_data_1 <- my_data %>% dplyr::select(heart, bp, smoke, age, male, exercise);

#Backward and forward Model selection
```

```r
backwards <- step(glm(factor(heart) ~ factor(bp) + factor(smoke) + age +
                        factor(male) + factor(exercise),
                      family = binomial,
                      data = my_data_1)); # Backwards selection is the default
```

```
## Start:  AIC=2005.14
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male) +
##     factor(exercise)
##
##                     Df Deviance    AIC
## - factor(exercise)  1   1994.8 2004.8
## <none>                  1993.1 2005.1
## - factor(smoke)     1   1998.8 2008.8
## - factor(male)      1   2009.1 2019.1
## - factor(bp)        1   2048.4 2058.4
## - age               1   2061.5 2071.5
##
## Step:  AIC=2004.85
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male)
##
##                 Df Deviance    AIC
## <none>              1994.8 2004.8
## - factor(smoke)  1   2000.3 2008.3
## - factor(male)   1   2009.9 2017.9
## - factor(bp)     1   2049.3 2057.3
## - age            1   2062.4 2070.4
```

```r
summary(backwards);
```

```
##
## Call:
## glm(formula = factor(heart) ~ factor(bp) + factor(smoke) + age +
##     factor(male), family = binomial, data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4266  -0.7832  -0.5533  -0.3433   2.3920
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.12025    0.13475 -15.735  < 2e-16 ***
## factor(bp)1     0.84386    0.11672   7.230 4.83e-13 ***
## factor(smoke)1  0.27623    0.11911   2.319 0.020382 *
## age             0.46642    0.05784   8.064 7.37e-16 ***
## factor(male)1   0.42974    0.11120   3.865 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2179.0  on 1978  degrees of freedom
## Residual deviance: 1994.8  on 1974  degrees of freedom
```

```
## AIC: 2004.8
##
## Number of Fisher Scoring iterations: 4
```

```
forward <- step(backwards,
                scope = list(lower = formula(backwards),
                upper = formula(backwards)),
                direction = "forward");
```

```
## Start:  AIC=2004.85
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male)
```

```
formula(forward);
```

```
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male)
```

```
model_1 <- glm(heart ~ bp + smoke + age + male,
               family = binomial, data = my_data_1);
summary(model_1);
```

```
##
## Call:
## glm(formula = heart ~ bp + smoke + age + male, family = binomial,
##     data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4266  -0.7832  -0.5533  -0.3433   2.3920
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.12025    0.13475 -15.735  < 2e-16 ***
## bp           0.84386    0.11672   7.230 4.83e-13 ***
## smoke        0.27623    0.11911   2.319 0.020382 *
## age          0.46642    0.05784   8.064 7.37e-16 ***
## male         0.42974    0.11120   3.865 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2179.0  on 1978  degrees of freedom
## Residual deviance: 1994.8  on 1974  degrees of freedom
## AIC: 2004.8
##
## Number of Fisher Scoring iterations: 4
```

```
# Check Correlation for Continuous variable
vif_1 <- car::vif(model_1);
vif_1;
```

```
##       bp    smoke      age     male
## 1.020299 1.003523 1.018526 1.001100
```

```r
#Step 2
#Backward and forward Model selection
backwards_2 <- step(glm(factor(heart) ~ factor(bp) + factor(smoke) + age +
                          factor(male) + factor(exercise) + smoke * male,
                        family = binomial,
                        data = my_data_1)); # Backwards selection is the default
```

```
## Start:  AIC=2001.02
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male) +
##     factor(exercise) + smoke * male
##
##
## Step:  AIC=2001.02
## factor(heart) ~ factor(bp) + factor(smoke) + age + factor(exercise) +
##     smoke + male + smoke:male
##
##
## Step:  AIC=2001.02
## factor(heart) ~ factor(bp) + age + factor(exercise) + smoke +
##     male + smoke:male
##
##                     Df Deviance    AIC
## - factor(exercise)  1   1988.9 2000.9
## <none>                  1987.0 2001.0
## - smoke:male        1   1993.1 2005.1
## - factor(bp)        1   2041.8 2053.8
## - age               1   2051.4 2063.4
##
## Step:  AIC=2000.9
## factor(heart) ~ factor(bp) + age + smoke + male + smoke:male
##
##              Df Deviance    AIC
## <none>           1988.9 2000.9
## - smoke:male  1   1994.8 2004.8
## - factor(bp)  1   2042.8 2052.8
## - age         1   2052.5 2062.5
```

```r
summary(backwards_2);
```

```
##
## Call:
## glm(formula = factor(heart) ~ factor(bp) + age + smoke + male +
##     smoke:male, family = binomial, data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4510  -0.7690  -0.5464  -0.3591   2.3716
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.91474    0.15413 -12.423  < 2e-16 ***
## factor(bp)1  0.84123    0.11692   7.195 6.25e-13 ***
## age          0.45430    0.05800   7.833 4.75e-15 ***
## smoke       -0.03096    0.17127  -0.181   0.8565
## male         0.03452    0.19639   0.176   0.8605
## smoke:male   0.58353    0.23933   2.438   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2179.0  on 1978  degrees of freedom
## Residual deviance: 1988.9  on 1973  degrees of freedom
## AIC: 2000.9
##
## Number of Fisher Scoring iterations: 4
```

```
forward_2 <- step(backwards_2,
                  scope = list(lower = formula(backwards_2),
                  upper = formula(backwards_2)),
                  direction = "forward");
```

```
## Start:  AIC=2000.9
## factor(heart) ~ factor(bp) + age + smoke + male + smoke:male
```

```
formula(forward_2);
```

```
## factor(heart) ~ factor(bp) + age + smoke + male + smoke:male
```

```
model_1_2 <- glm(factor(heart) ~ factor(bp) + factor(smoke) + age +
                   factor(male) + factor(smoke) * factor(male),
                 family = binomial, data = my_data_1);
summary(model_1_2);
```

```
##
## Call:
## glm(formula = factor(heart) ~ factor(bp) + factor(smoke) + age +
##     factor(male) + factor(smoke) * factor(male), family = binomial,
##     data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4510  -0.7690  -0.5464  -0.3591   2.3716
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.91474    0.15413 -12.423  < 2e-16 ***
## factor(bp)1                     0.84123    0.11692   7.195 6.25e-13 ***
## factor(smoke)1                 -0.03096    0.17127  -0.181   0.8565
## age                             0.45430    0.05800   7.833 4.75e-15 ***
## factor(male)1                   0.03452    0.19639   0.176   0.8605
## factor(smoke)1:factor(male)1    0.58353    0.23933   2.438   0.0148 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2179.0  on 1978  degrees of freedom
## Residual deviance: 1988.9  on 1973  degrees of freedom
## AIC: 2000.9
##
## Number of Fisher Scoring iterations: 4
```

```
#Likelihood Ratio Test
lrtest(model_1, model_1_2);
```

```
## Likelihood ratio test
##
## Model 1: heart ~ bp + smoke + age + male
## Model 2: factor(heart) ~ factor(bp) + factor(smoke) + age + factor(male) +
##     factor(smoke) * factor(male)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -997.42
## 2   6 -994.45  1 5.9481    0.01473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Step 3
## Odds Ratios and Confidence Intervals
exp(cbind(OR = coef(model_1), confint(model_1)));
```

```
## Waiting for profiling to be done...
```

```
##                    OR      2.5 %    97.5 %
## (Intercept) 0.1200015 0.09169282 0.1555455
## bp          2.3253222 1.85276921 2.9284631
## smoke       1.3181546 1.04531760 1.6677834
## age         1.5942818 1.42440626 1.7871123
## male        1.5368637 1.23656143 1.9125277
```

```
exp(cbind(OR = coef(model_1_2), confint(model_1_2)));
```

```
## Waiting for profiling to be done...
```

```
##                                       OR     2.5 %    97.5 %
## (Intercept)                   0.1473805 0.1081613 0.1980208
## factor(bp)1                   2.3192170 1.8471431 2.9219117
## factor(smoke)1                0.9695133 0.6945899 1.3602836
## age                           1.5750709 1.4067971 1.7661096
## factor(male)1                 1.0351266 0.7037758 1.5213401
## factor(smoke)1:factor(male)1 1.7923560 1.1214709 2.8676706
```

```r
pander(c("Model_1 AIC" = AIC(model_1), "Model_1_2 AIC" = AIC(model_1_2),
         "Model_1 BIC" = BIC(model_1), "Model_1_2 BIC" = BIC(model_1_2)));
```

Model_1 AIC	Model_1_2 AIC	Model_1 BIC	Model_1_2 BIC
2005	2001	2033	2034

```r
## ANOVA Comparison
anova(model_1, model_1_2)
```

```
## Warning in anova.glmlist(c(list(object), dotargs), dispersion = dispersion, :
## models with response '"factor(heart)"' removed because response differs from
## model 1

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: heart
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev
## NULL                  1978     2179.0
## bp     1   90.835      1977     2088.1
## smoke  1    7.170      1976     2080.9
## age    1   71.047      1975     2009.9
## male   1   15.057      1974     1994.8
```

```r
## 1- chisq probability
### Basic Model
pander(c(1 - pchisq(model_1$deviance, model_1$df.residual), 1 - pchisq(model_1_2$deviance, model_1_2$df
```

0.3665 and *0.3963*

```r
## Hosmer-Lemeshow Test
### Basic Model
ht_model_1 <- hoslem.test(model_1$y, fitted(model_1), g = 10)
ht_model_1
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_1$y, fitted(model_1)
## X-squared = 22.798, df = 8, p-value = 0.003633
```

```r
### Interaction Model
ht_model_1_2 <- hoslem.test(model_1_2$y, fitted(model_1_2), g = 10)
ht_model_1_2
```

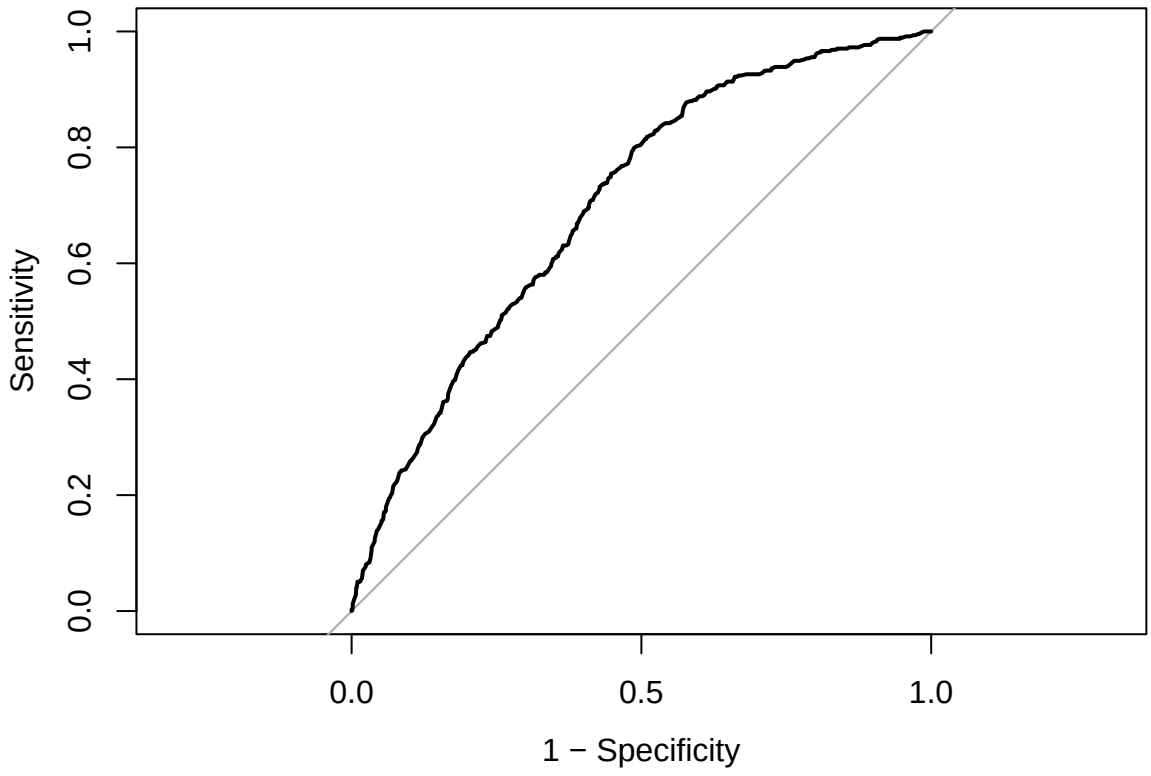
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_1_2$y, fitted(model_1_2)
## X-squared = 9.6127, df = 8, p-value = 0.2933
```

```
# Plot ROC curves and AUC
rocplot1 <- roc(heart~fitted(model_1), data=my_data_1);
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

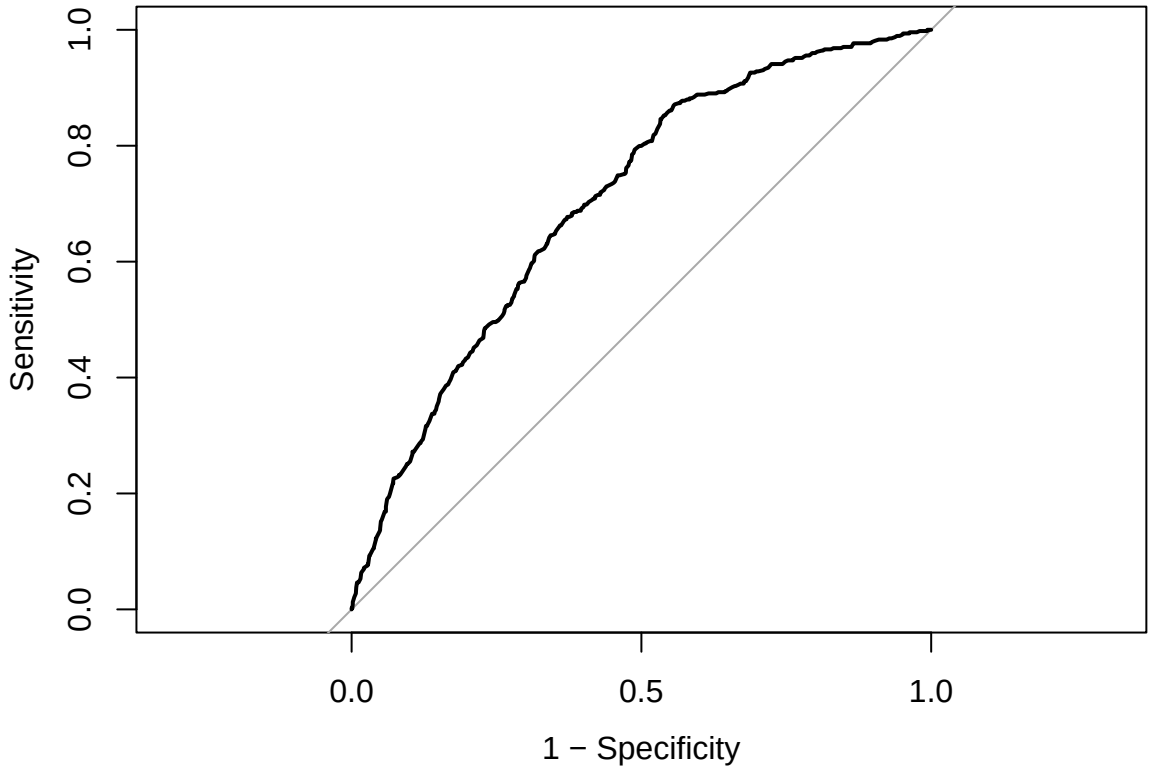
```
plot.roc(rocplot1, legacy.axes=TRUE);
```




```
rocplot2 <- roc(heart~fitted(model_1_2), data=my_data_1);
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot.roc(rocplot2, legacy.axes=TRUE);
```




```
pander(c("Model_1 AUC" = auc(my_data_1$heart, model_1$fitted.values),
         "Model_1_2 AUC" = auc(my_data_1$heart, model_1_2$fitted.values)));
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

Model_1 AUC	Model_1_2 AUC
0.701	0.7048

Case2 For the second case study, build a predictive model using a self-rated health category for individuals in their midlife as an outcome variable. Select the correct model based on the distribution of the outcome variable. In step 1, you will fit and conduct the main effects model with the variables depress, alcage, cigage, age, and bp. In step 2, you will test any relevant assumptions. In step 3 [2 bonus points], predict probabilities for those with diagnosed high blood pressure if (1) age = 64 and depress = 0, and (2) if age = 64 and depress = 1. Report the difference in probabilities per group (i.e., excellent, good, fair, poor).

```r
# this is a cumulative logit model for ordinal response.
my_data_2 <- my_data %>% dplyr::select(health, depress, alcage, cigage, age, bp);
#Step 1
model_2 <- vglm(factor(health) ~ factor(depress) + alcage + cigage + age + factor(bp),
                family = cumulative(parallel=TRUE), data = my_data_2);
summary(model_2);
```

```
##
## Call:
## vglm(formula = factor(health) ~ factor(depress) + alcage + cigage +
##     age + factor(bp), family = cumulative(parallel = TRUE), data = my_data_2)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept):1      0.50482    0.06716   7.517 5.60e-14 ***
## (Intercept):2      2.18173    0.08398  25.978  < 2e-16 ***
## (Intercept):3      3.66766    0.12178  30.117  < 2e-16 ***
## factor(depress)1 -1.20181    0.15437  -7.785 6.97e-15 ***
## alcage            0.01399    0.04565   0.306    0.759
## cigage           -0.02180    0.04516  -0.483    0.629
## age              -0.06050    0.04588  -1.319    0.187
## factor(bp)1      -0.94553    0.09015 -10.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 4305.659 on 5929 degrees of freedom
##
## Log-likelihood: -2152.83 on 5929 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):3'
##
##
## Exponentiated coefficients:
## factor(depress)1           alcage           cigage              age
##        0.3006510        1.0140846        0.9784385        0.9412895
##      factor(bp)1
##        0.3884738
```

```r
vif_model <- lm(health ~ depress + alcage + cigage + age + bp, data = my_data_2);
vif_2 <- vif(vif_model);
vif_2;
```

```
##  depress   alcage   cigage      age       bp
## 1.009765 1.149760 1.116946 1.130714 1.057105
```

```r
#Step 2
## Relative Odd Ratio
(fit.rr.2.1 <- exp(coefficients(model_2)))
```

```
##    (Intercept):1    (Intercept):2    (Intercept):3 factor(depress)1
##        1.6566865        8.8615813       39.1600735        0.3006510
##           alcage           cigage              age      factor(bp)1
##        1.0140846        0.9784385        0.9412895        0.3884738
```

```r
## Consider only the significant variables in the model
table(my_data_2$health, my_data_2$depress)
```

```
##
##       0   1
##   1 913  50
##   2 619  40
##   3 212  45
##   4  77  23
```

```r
table(my_data_2$health, my_data_2$bp)
```

```
##
##       0   1
##   1 581 382
##   2 281 378
##   3  85 172
##   4  19  81
```

```r
#For depress
odd.ratio.dp1 <- ((77+212+619)*50)/((23+45+40)*913)
odd.ratio.dp2 <- ((77+212)*(40+50))/((23+45)*(619+913))
odd.ratio.dp3 <- (77*(50+40+45))/(23*(212+619+913))
odd.ratio.dp <- odd.ratio.dp1 + odd.ratio.dp2 + odd.ratio.dp3

#For bp
odd.ratio.bp1 <- ((19+85+281)*382)/((81+172+378)*581)
odd.ratio.bp2 <- ((19+85)*(382+378))/((81+172)*(581+281))
odd.ratio.bp3 <- (19*(382+378+172))/(81*(581+281+85))
odd.ratio.bp <- odd.ratio.bp1 + odd.ratio.bp2 + odd.ratio.bp3

odd.ratio.case2 <- data.frame(
  'Variable' = c('odd.ratio.dp1', 'odd.ratio.dp2', 'odd.ratio.dp3', 'odd.ratio.dp',
                 'odd.ratio.bp1', 'odd.ratio.bp2', 'odd.ratio.bp3', 'odd.ratio.bp'),
  'Value' = c(odd.ratio.dp1, odd.ratio.dp2, odd.ratio.dp3, odd.ratio.dp,
              odd.ratio.bp1, odd.ratio.bp2, odd.ratio.bp3, odd.ratio.bp)
)
pander(odd.ratio.case2)
```

Variable	Value
odd.ratio.dp1	0.4604
odd.ratio.dp2	0.2497
odd.ratio.dp3	0.2591
odd.ratio.dp	0.9693
odd.ratio.bp1	0.4012
odd.ratio.bp2	0.3624
odd.ratio.bp3	0.2309
odd.ratio.bp	0.9944

```r
#Step 3
#my_data_2_3 <- data.frame(depress = c(0, 1), alcage = mean(my_data$alcage, na.rm = TRUE),
                      # cigage = mean(my_data$cigage, na.rm = TRUE), age = 64, bp = 1)
result <- predict(model_2, newdata = my_data_2_3, type = "response");
not_equal_to_1_indices <- matrix(nrow = nrow(result))
for (i in 1:nrow(result)) {
  if (sum(result[i,]) != 1) {
    not_equal_to_1_indices[i,] <- i
  }
}
not_equal_to_1_indices
```

```
##       [,1]
##  [1,]   NA
##  [2,]   NA
##  [3,]   NA
##  [4,]   NA
##  [5,]   NA
##  [6,]   NA
##  [7,]   NA
##  [8,]   NA
##  [9,]   NA
## [10,]   NA
## [11,]   NA
## [12,]   NA
## [13,]   NA
## [14,]   NA
## [15,]   NA
## [16,]   NA
## [17,]   NA
## [18,]   NA
## [19,]   NA
## [20,]   NA
## [21,]   NA
## [22,]   NA
## [23,]   NA
## [24,]   NA
## [25,]   NA
## [26,]   NA
## [27,]   27
## [28,]   NA
## [29,]   NA
```

```
## [30,]   NA
## [31,]   NA
## [32,]   32
```