HW3 Q2

Karen Hong and Juyi Yang

2024-12-06

1.

```
df <-  read.csv("~/Downloads/diabetic.csv")
# a)
df$TotalRiskFactors = factor(ifelse(df$TotalRiskFactors == 0, "none",
                                    ifelse(df$TotalRiskFactors == 1, "one", "2 or more")), levels = c("none"

# b)
df$Diabetes.new = ifelse(df$Diabetes.new == 0, "no", "yes")

# c)
df$SmokingStatus_NISTCode = factor(df$SmokingStatus_NISTCode,
                                         levels = c("FALSE", "FORMER", "TRUE"))

# d)
df$age.new = factor(df$age.new, levels = c("low", "medium", "high"))
```

e.

```
age <- table(df$TotalRiskFactors,df$age.new)
prop.table(age, margin = 1)
```

```
##
##                    low    medium      high
##   none      0.4750450 0.3149730 0.2099820
##   one       0.2116767 0.3678458 0.4204774
##   2 or more 0.2179912 0.3476821 0.4343267
```

It seems like the no risk factos group has the highest proportion of the lowest age group while one and 2 or more risk factors have a much higher proportion of the high age group

```
d <- table(df$Diabetes.new,df$age.new)
prop.table(d, margin = 1)
```

```
##
##              low     medium       high
##   no  0.38655988 0.33525216 0.27818796
##   yes 0.08877434 0.36139748 0.54982818
```

Meanwhile those with diabetes tend to have a lower proportion of the low age group and a very high proportion of the high age group while it's spread more evenly across those with no diabetes.

```
s <- table(df$SmokingStatus_NISTCode,df$age.new)
prop.table(s, margin = 1)
```

```
##
##                low    medium      high
##   FALSE  0.2303160 0.3176219 0.4520621
##   FORMER 0.3253925 0.3107742 0.3638332
##   TRUE   0.3460899 0.4326123 0.2212978
```

Those who aren't smokers have the highest proportion of the high age group and this would make sense as thoe who don't smoke tend to live longer. Forformer smokers, it seems to be roughly evenly split across all groups while the smokers tend to have the highest proportion of low and medium age groups which is again not surprising as not many smokers live very long.

```
h <- table(df$HypertensionDX,df$age.new)
prop.table(h, margin = 1)
```

```
##
##             low    medium      high
##   no  0.4870712 0.3208443 0.1920844
##   yes 0.1172261 0.3667325 0.5160415
```

We can see that those who don't have hypertension have the highest proportion of the low age group while those with hypertension has the highest proportion of the high age group.

2.

```
valid_columns = c('TotalRiskFactors', 'Diabetes.new', 'HypertensionDX', 'SmokingStatus_NISTCode')
valid_columns = valid_columns[valid_columns %in% colnames(df)]

if (length(valid_columns) == 0) {
  stop('None of the specified predictor columns exist in the dataset.')
}

levels_info = sapply(df[, valid_columns, drop = FALSE], function(x) length(unique(x)))

print(levels_info)
```

```
##       TotalRiskFactors           Diabetes.new         HypertensionDX
##                      3                      2                      2
## SmokingStatus_NISTCode
##                      4
```

```
for (col in names(levels_info)) {
  if (levels_info[col] == 1) {
    cat(sprintf( col))
    df[[col]] = NULL
  }
```

```
}
library(nnet)
predictors = valid_columns[valid_columns %in% colnames(df)]
formula = as.formula(paste('age.new ~', paste(predictors, collapse = ' + ')))
model = multinom(formula, data = df)
```

```
## # weights:  24 (14 variable)
## initial  value 4740.512026
## iter  10 value 4260.991992
## iter  20 value 4108.350866
## final  value 4108.350692
## converged
```

3.

```
(coef <- summary(model)$coefficients)
```

```
##        (Intercept) TotalRiskFactorsone TotalRiskFactors2 or more
## medium  -0.5158245           0.7201609                 0.6742832
## high    -0.8435857           0.9580151                 1.0635043
##        Diabetes.newyes HypertensionDXyes SmokingStatus_NISTCodeFORMER
## medium       0.8848004          1.336294                   -0.3895716
## high         1.1620844          2.064695                   -0.6001646
##        SmokingStatus_NISTCodeTRUE
## medium                 -0.2048692
## high                   -1.2855500
```

```
(intv <- exp(confint(model)))
```

```
## , , medium
##
##                                    2.5 %    97.5 %
## (Intercept)                    0.5105510 0.6981061
## TotalRiskFactorsone            1.7057859 2.4751373
## TotalRiskFactors2 or more      1.5425623 2.4970789
## Diabetes.newyes                1.8100098 3.2422532
## HypertensionDXyes              3.1528573 4.5918309
## SmokingStatus_NISTCodeFORMER 0.5665288 0.8098423
## SmokingStatus_NISTCodeTRUE   0.6427466 1.0327926
##
## , , high
##
##                                    2.5 %    97.5 %
## (Intercept)                    0.3639352 0.5084482
## TotalRiskFactorsone            2.1430714 3.1701856
## TotalRiskFactors2 or more      2.2655634 3.7031547
## Diabetes.newyes                2.3969790 4.2629421
## HypertensionDXyes              6.5113559 9.5433334
## SmokingStatus_NISTCodeFORMER 0.4571262 0.6586696
## SmokingStatus_NISTCodeTRUE   0.2093024 0.3652676
```

```
z <-summary(model)$coefficients/summary(model)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1))*2

t1 <- cbind("log odds" = coef["medium",],odds = exp(coef["medium",]),intv[,,"medium"],"p-value" = p['me
t2 <- cbind("log odds" = coef["high",],odds = exp(coef["high",]),intv[,,"high"],"p-value" =p['high',])[
```

4. Medium

```
as.data.frame(t1)
```

```
##                                 log odds      odds     2.5 %    97.5 %
## TotalRiskFactorsone            0.7201609 2.0547638 1.7057859 2.4751373
## TotalRiskFactors2 or more      0.6742832 1.9626257 1.5425623 2.4970789
## Diabetes.newyes                0.8848004 2.4225008 1.8100098 3.2422532
## HypertensionDXyes              1.3362940 3.8049162 3.1528573 4.5918309
## SmokingStatus_NISTCodeFORMER -0.3895716 0.6773470 0.5665288 0.8098423
## SmokingStatus_NISTCodeTRUE   -0.2048692 0.8147539 0.6427466 1.0327926
##                                    p-value
## TotalRiskFactorsone           3.375078e-14
## TotalRiskFactors2 or more     4.079465e-08
## Diabetes.newyes               2.684713e-09
## HypertensionDXyes             0.000000e+00
## SmokingStatus_NISTCodeFORMER 1.921347e-05
## SmokingStatus_NISTCodeTRUE    9.040271e-02
```

High

```
as.data.frame(t2)
```

```
##                                 log odds      odds     2.5 %    97.5 %
## TotalRiskFactorsone            0.9580151 2.6065176 2.1430714 3.1701856
## TotalRiskFactors2 or more      1.0635043 2.8965034 2.2655634 3.7031547
## Diabetes.newyes                1.1620844 3.1965893 2.3969790 4.2629421
## HypertensionDXyes              2.0646953 7.8828954 6.5113559 9.5433334
## SmokingStatus_NISTCodeFORMER -0.6001646 0.5487213 0.4571262 0.6586696
## SmokingStatus_NISTCodeTRUE   -1.2855500 0.2764985 0.2093024 0.3652676
##                                    p-value
## TotalRiskFactorsone           0.000000e+00
## TotalRiskFactors2 or more     0.000000e+00
## Diabetes.newyes               2.442491e-15
## HypertensionDXyes             0.000000e+00
## SmokingStatus_NISTCodeFORMER 1.188052e-10
## SmokingStatus_NISTCodeTRUE    0.000000e+00
```

5.

a) The odds of risk factor Interpretation: The rows TotalRiskFactorsone and TotalRiskFactors2 or more provide the odds for individuals with one risk factor and two or more risk factors, respectively.

The odds of having a high age compared to low age is

2.61 times higher for one risk factor compared to no risk factor 2.9 times higher for 2+ risk factors compared to no risk factor

b) The odds of diabetic type II Interpretation: The row Diabetes.newyes refers to individuals with diabetes type II. The odds of diabetes type II being associated with the outcome of interest are 3.20. This means that the odds of having a high age compared to low age is 3.2 times higher for those with diabetes compared to those who don't

c) The odds of smoker Interpretation: The rows SmokingStatus_NISTCodeFORMER and SmokingStatus_NISTCodeTRUE provide the odds for former smokers and current smokers, respectively. This means that the odds of having a high age compared to low age is

45% lower for for former smokers compared to non-smokers and 72% lower for smokers compared to non-smokers

d) The 95% confidence interval for the odds of diabetic type II Interpretation: The row Diabetes.newyes has a 95% confidence interval of (2.40, 4.26). This means that we are 95% confident that the true odds of high vs. low for individuals with diabetes type II as opposed to not having it lie between 2.40 and 4.26. Since this interval does not include 1, the association is statistically significant.

e) The 95% confidence interval for smoker Interpretation: For former smokers (SmokingStatus_NISTCodeFORMER), the 95% confidence interval is (0.46, 0.66). This indicates a statistically significant reduction in the odds of high vs low compared to non-smokers. For current smokers (SmokingStatus_NISTCodeTRUE), the 95% confidence interval is (0.21, 0.37). This also shows a statistically significant reduction in the odds of high vs low compared to non-smokers.

6.

$log(\frac{P(High)}{P(Low)})$=-.84 +.96(1 Risk Factor) + 1.06(2+ Risk Factors)+1.16(Diabetes)+2.06(Hypertension) -.6(Former Smoker)-1.29(Smoker)

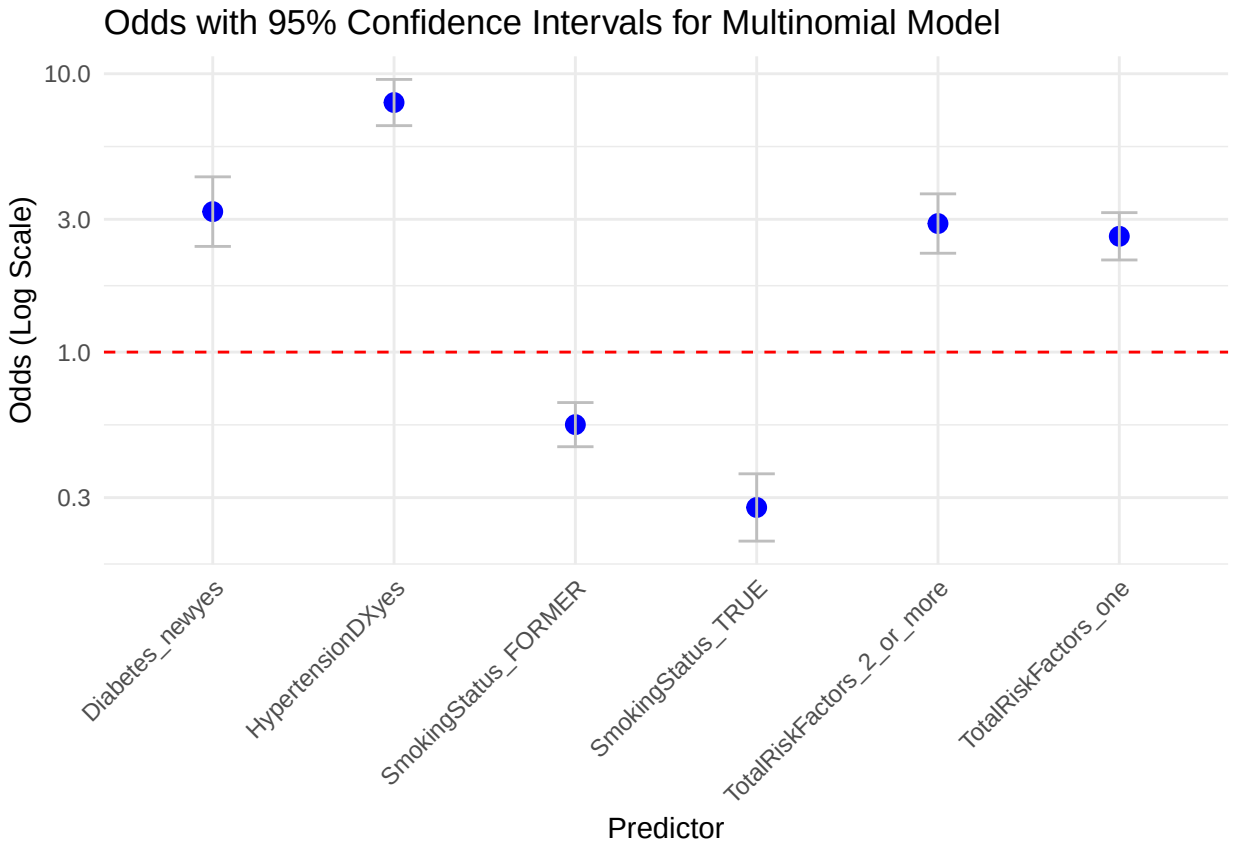
7.

```r
#7
# Load necessary library
library(ggplot2)

# Data from the table
data <- data.frame(
  Predictor = c(
    "TotalRiskFactors_one",
    "TotalRiskFactors_2_or_more",
    "Diabetes_newyes",
    "HypertensionDXyes",
    "SmokingStatus_FORMER",
    "SmokingStatus_TRUE"
  ),
  Odds = c(2.6065176, 2.8965034, 3.1965893, 7.8828954, 0.5487213, 0.2764985),
  Lower_CI = c(2.1430714, 2.2655634, 2.3969790, 6.5113559, 0.4571262, 0.2093024),
  Upper_CI = c(3.1701856, 3.7031547, 4.2629421, 9.5433334, 0.6586696, 0.3652676)
)

# Plot using ggplot2
ggplot(data, aes(x = Predictor, y = Odds)) +
  geom_point(size = 3, color = "blue") +
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI), width = 0.2, color = "gray") +
  geom_hline(yintercept = 1, linetype = "dashed", color = "red") +
```

```r
  scale_y_log10() +  # Use log scale for odds
  labs(
    title = "Odds with 95% Confidence Intervals for Multinomial Model",
    x = "Predictor",
    y = "Odds (Log Scale)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



#8 Based on the plot of odds with 95% confidence intervals, the following summarizes the effects of predictors on the outcome (high vs. low age):

Diabetes_newyes: Being diabetic is associated with an increased likelihood of the outcome in the high-age group compared to the low-age group.

HypertensionDXyes: A diagnosis of hypertension is strongly associated with an increased likelihood of the outcome in the high-age group compared to the low-age group.

SmokingStatus_FORMER: Former smoking status is associated with a decreased likelihood of the outcome in the high-age group compared to the low-age group.

SmokingStatus_TRUE: Current smoking status is associated with a further decreased likelihood of the outcome in the high-age group compared to the low-age group.

TotalRiskFactors_one: Having one risk factor is associated with a modestly increased likelihood of the outcome in the high-age group compared to the low-age group.

TotalRiskFactors_2_or_more: Having two or more risk factors is associated with a slightly greater increase in the likelihood of the outcome compared to having one risk factor in the high-age group.

The error bars represent the uncertainty (95% confidence intervals) around these effects. Predictors whose confidence intervals do not cross the red reference line (odds = 1) are statistically significant contributors to the outcome. This indicates that the predictors collectively differentiate between high and low age groups with respect to the outcome of interest.

```
#9
# Load necessary library
library(caret)
```

```
## Loading required package: lattice
```

```
predicted <- predict(model,df)

# Create a confusion matrix
confusion_matrix <- confusionMatrix(predicted, df$age.new)

# Print the confusion matrix
print(confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  low medium high
##     low     916    513  282
##     medium  126    260  226
##     high    197    654 1141
##
## Overall Statistics
##
##                Accuracy : 0.537
##                  95% CI : (0.5219, 0.5519)
##     No Information Rate : 0.3822
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3014
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: low Class: medium Class: high
## Sensitivity              0.7393       0.18220      0.6919
## Specificity              0.7415       0.87812      0.6808
## Pos Pred Value           0.5354       0.42484      0.5728
## Neg Pred Value           0.8760       0.68485      0.7813
## Prevalence               0.2871       0.33071      0.3822
## Detection Rate           0.2123       0.06025      0.2644
## Detection Prevalence     0.3965       0.14183      0.4616
## Balanced Accuracy        0.7404       0.53016      0.6864
```

```
# Extract the accuracy
accuracy <- confusion_matrix$overall['Accuracy']
print(paste("Accuracy:", accuracy))
```

```
## [1] "Accuracy: 0.536964078794901"
```

#10 === Model Performance Analysis ===

Areas where the model performed well: Accuracy:

The overall accuracy of the model is 53.7%, meaning the model correctly predicted the class labels for 53.7% of the observations. While not perfect, this is higher than the no-information rate (NIR) of 38.2%, indicating the model performs better than random guessing. Sensitivity (Class: High & Low):

The sensitivity for the High and Low classes is 69.2% and 73.9% respectively. This means the model correctly identified 69.2% and 73.9% of the actual "High" and "Low" cases respectively. This is a reasonably strong performance in terms of capturing these classes. Specificity (Class: High):

The sensitivity for the Medium class is 18.2%, meaning the model only correctly identifies 18.2% of the actual "Medium" cases which is very low. This indicates room for improvement in identifying this class. Specificity (Class: Medium):

The specificity for the High class is 68.1%, meaning that the model misclassifies 31.9% of the non-"High" observations as the "High" category which is the lowest among all the classes. Positive Predictive Value (Class: High):

The specificity for the Medium class is 87.8%, meaning that the model misclassifies 12.2% of the non-"Medium" cases as "Medium." This is relatively low compared to the "High" and "Low" classes. Balanced Accuracy (Class: Medium):

The specificity for the Low class is 74.2%, meaning that 15.8% of non-"Low" cases are incorrectly classified as "Low." which isn't too bad.

Summary: The model performed best for the Low class, with strong sensitivity, specificity. The model had moderate performance for the High class, with reasonable sensitivity but slightly lower specificity. The model performed poorly for the Medium class, struggling with sensitivity, indicating that it had difficulty distinguishing this class from the others.