Question 4 Syahid

Ahmad Firdaus Mohamed

2024-02-10

library(haven)  
library(readxl)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(broom)  
library(gtsummary)  
library(ggplot2)  
library(lattice)  
library(caret)

##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(pROC)

## Type 'citation("pROC")' for a citation.  
##   
## Attaching package: 'pROC'  
##   
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(ResourceSelection)

## ResourceSelection 0.3-6 2023-06-27

library(performance)  
#library(LogisticDx)  
library(largesamplehl)

Data1 <- read\_excel('peptic\_ulcer.xlsx')  
Data1 <- Data1 %>%  
 mutate(across(where(is.character),as\_factor))

summary(Data1)

## age gender epigastric\_pain vomiting nausea fever   
## Min. :19.00 male :96 yes:116 no :78 no :113 no :96   
## 1st Qu.:49.00 female:25 no : 5 yes:43 yes: 8 yes:25   
## Median :64.00   
## Mean :60.43   
## 3rd Qu.:75.00   
## Max. :97.00   
##   
## diarrhea malena onset\_more\_24\_hrs NSAIDS septic\_shock previous\_OGDS  
## no :115 no :117 no :48 no :86 no :106 no :111   
## yes: 6 yes: 4 yes:73 yes:35 yes: 15 yes: 10   
##   
##   
##   
##   
##   
## ASA systolic diastolic inotropes pulse   
## Min. :1.000 Min. : 67.0 Min. : 38.00 no :110 Min. : 55.00   
## 1st Qu.:1.000 1st Qu.:112.0 1st Qu.: 63.00 yes: 11 1st Qu.: 82.00   
## Median :1.000 Median :128.0 Median : 71.00 Median : 95.00   
## Mean :1.545 Mean :128.6 Mean : 72.07 Mean : 94.69   
## 3rd Qu.:2.000 3rd Qu.:143.0 3rd Qu.: 81.00 3rd Qu.:105.00   
## Max. :3.000 Max. :197.0 Max. :116.00 Max. :155.00   
##   
## tenderness guarding hemoglobin twc platelet   
## generalized:84 yes:86 Min. : 3.30 Min. : 2.00 Min. : 11.0   
## localized :37 no :35 1st Qu.:10.00 1st Qu.: 9.00 1st Qu.:224.0   
## Median :12.00 Median :12.00 Median :308.0   
## Mean :12.32 Mean :13.03 Mean :314.4   
## 3rd Qu.:15.00 3rd Qu.:16.00 3rd Qu.:392.0   
## Max. :19.40 Max. :37.00 Max. :798.0   
##   
## creatinine albumin PULP admission\_to\_op\_hrs  
## Min. : 37.0 38 : 9 Min. :0.000 Min. : 1.00   
## 1st Qu.: 78.0 35 : 8 1st Qu.:2.000 1st Qu.: 5.00   
## Median :113.0 37 : 7 Median :3.000 Median : 8.00   
## Mean :147.4 27 : 6 Mean :3.529 Mean :10.07   
## 3rd Qu.:168.0 30 : 6 3rd Qu.:5.000 3rd Qu.:12.00   
## Max. :843.0 40 : 6 Max. :9.000 Max. :72.00   
## (Other):79   
## perforation degree\_perforation side\_perforation ICU SSSI   
## Min. :0.300 small :75 distal stomach:69 no :48 no :110   
## 1st Qu.:0.500 large :26 duodenum :37 yes:73 yes: 11   
## Median :1.000 moderate:20 prepyloric : 6   
## Mean :1.225 body stomach : 8   
## 3rd Qu.:1.500 body : 1   
## Max. :5.000   
##   
## anast\_leak sepsis outcome   
## no :113 no :77 alive:83   
## yes: 8 yes:44 dead :38   
##   
##   
##   
##   
##

A dataset named Q4.csv has six (6) variables. (? Dataset <https://github.com/drkamarul/data-for-RMed/blob/main/data/peptic_ulcer.xlsx>) (SYAHID)

1. Outcome: alive or dead (the outcome variable)

ii.Sepsis: yes or no

1. Creatinine: in mmol/L
2. Age: in years old
3. Perforation: size of perforation in cm
4. Hemoglobin: in mg/dL

The **binary logistic regression analysis** was chosen to assess the relationship between the outcome and the rest of the variables.

## (a) Justify the choice of the analysis. (4 marks)

The dependant variable is dichotomous and the independent variables are numerical and categorical variables

State the assumptions for the analysis. (2 marks)

there is linearity of logit with the covariate

the outcome variable is binomial and has bernouli distribution

absence of multicollinearity

fixed errors (measured without error)

independe

independent variable

## (b) Generate a saturated model and name it as Model\_Outcome. Present Model\_Outcome in a table. (5 marks)

Model\_Outcome <- glm (outcome ~ sepsis + creatinine + age + perforation + hemoglobin, family = binomial(link='logit'), data = Data1)

summary(Model\_Outcome)

##   
## Call:  
## glm(formula = outcome ~ sepsis + creatinine + age + perforation +   
## hemoglobin, family = binomial(link = "logit"), data = Data1)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.182999 2.768177 -3.317 0.000909 \*\*\*  
## sepsisyes 4.816703 0.921962 5.224 1.75e-07 \*\*\*  
## creatinine 0.009775 0.003635 2.689 0.007157 \*\*   
## age 0.005208 0.021257 0.245 0.806448   
## perforation 0.854089 0.355906 2.400 0.016406 \*   
## hemoglobin 0.240894 0.119680 2.013 0.044134 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 150.597 on 120 degrees of freedom  
## Residual deviance: 58.392 on 115 degrees of freedom  
## AIC: 70.392  
##   
## Number of Fisher Scoring iterations: 6

TableMLogR <- tbl\_regression(Model\_Outcome)  
TableMLogR

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **log(OR)**1 | **95% CI**1 | **p-value** |
| --- | --- | --- | --- |
| sepsis |  |  |  |
| no | — | — |  |
| yes | 4.8 | 3.2, 7.0 | <0.001 |
| creatinine | 0.01 | 0.00, 0.02 | 0.007 |
| age | 0.01 | -0.04, 0.05 | 0.8 |
| perforation | 0.85 | 0.18, 1.6 | 0.016 |
| hemoglobin | 0.24 | 0.02, 0.49 | 0.044 |
| 1OR = Odds Ratio, CI = Confidence Interval | | | |

## (c)Using Model\_Outcome, compose an interaction between creatinine and sepsis. Name it as Model\_Outcome\_ia. Present Model\_Outcome\_ia in a table. (5 marks)

Model\_Outcome\_ia <- glm(formula = outcome ~ sepsis + creatinine + age + perforation + hemoglobin + creatinine:sepsis , family = binomial(link = "logit"), data = Data1)

summary(Model\_Outcome\_ia)

##   
## Call:  
## glm(formula = outcome ~ sepsis + creatinine + age + perforation +   
## hemoglobin + creatinine:sepsis, family = binomial(link = "logit"),   
## data = Data1)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -12.12348 3.85631 -3.144 0.00167 \*\*  
## sepsisyes 8.62090 2.83350 3.042 0.00235 \*\*  
## creatinine 0.02536 0.01463 1.733 0.08301 .   
## age 0.00541 0.02161 0.250 0.80231   
## perforation 0.87770 0.36225 2.423 0.01540 \*   
## hemoglobin 0.25045 0.11773 2.127 0.03340 \*   
## sepsisyes:creatinine -0.02282 0.01522 -1.500 0.13367   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 150.597 on 120 degrees of freedom  
## Residual deviance: 52.702 on 114 degrees of freedom  
## AIC: 66.702  
##   
## Number of Fisher Scoring iterations: 8

TableMLogR.ia <- tbl\_regression(Model\_Outcome\_ia)  
TableMLogR.ia

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **log(OR)**1 | **95% CI**1 | **p-value** |
| --- | --- | --- | --- |
| sepsis |  |  |  |
| no | — | — |  |
| yes | 8.6 | 4.7, 16 | 0.002 |
| creatinine | 0.03 | 0.01, 0.06 | 0.083 |
| age | 0.01 | -0.04, 0.05 | 0.8 |
| perforation | 0.88 | 0.19, 1.7 | 0.015 |
| hemoglobin | 0.25 | 0.03, 0.50 | 0.033 |
| sepsis \* creatinine |  |  |  |
| yes \* creatinine | -0.02 | -0.06, 0.00 | 0.13 |
| 1OR = Odds Ratio, CI = Confidence Interval | | | |

## (d) For Model\_Outcome\_ia, write:

The logistic model. (2 marks)

E(Y) = P = (Exp^(B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6) / (1 + Exp ^ (B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6))

The logit equation. (2 marks)

Logit (Y) = -12.1234 + 8.6209 (sepsisyes) + 0.0253(creatinine) + 0.0054 (age) + 0.8777 (perforation) + 0.25045 (hemoglobin) - 0.0228 (sepsisyes\*creatinine)

## (e) Using Model\_Outcome\_ia, generate the predicted values for Observation 1. Show your manual calculation. (4 marks)

augment(Model\_Outcome\_ia)

## # A tibble: 121 × 12  
## outcome sepsis creatinine age perforation hemoglobin .fitted .resid  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 alive no 135 42 0.5 18 -3.53 -0.241   
## 2 alive no 66 66 1 12 -6.21 -0.0634  
## 3 alive no 80 67 0.5 12 -6.29 -0.0610  
## 4 alive no 64 19 0.5 12 -6.95 -0.0437  
## 5 alive no 137 77 1 11 -4.60 -0.141   
## 6 alive no 102 39 1 18 -3.94 -0.196   
## 7 dead yes 69 62 3 8.1 1.67 0.587   
## 8 alive no 92 71 1.5 13.3 -4.76 -0.131   
## 9 alive no 94 69 0.5 11.8 -5.97 -0.0714  
## 10 alive no 104 97 1.5 10 -5.14 -0.108   
## # ℹ 111 more rows  
## # ℹ 4 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>

-12.1234 + (8.6209\*0) + (0.0253\*135) + (0.0054\*42) + (0.8777\*0.5) + (0.25045\*18) - (0.0228\*0\*135)

## [1] -3.53415

## Model Assessment

### Classification Table

fit.m <- augment(Model\_Outcome\_ia, type.predict = 'response') %>%  
 mutate(pred.class=factor(ifelse(.fitted>0.5,'dead','alive')))

confusionMatrix(fit.m$outcome,fit.m$pred.class)

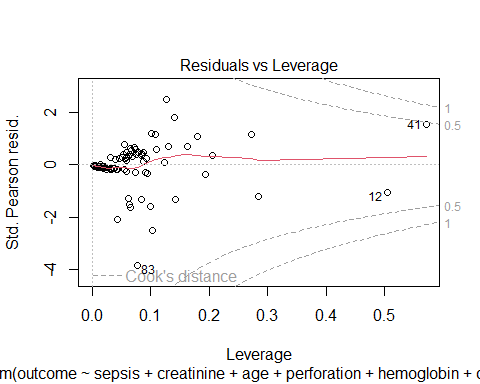
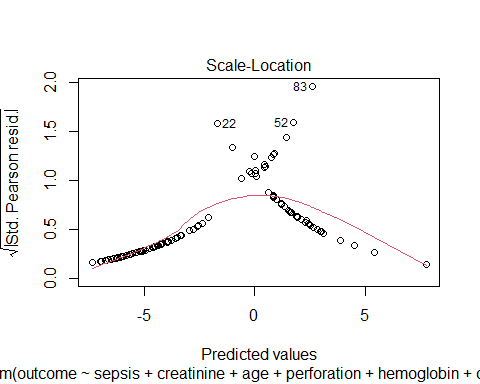
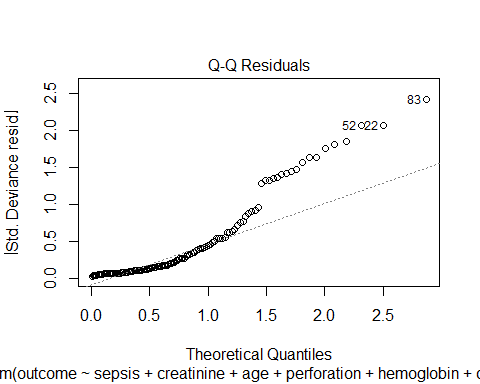
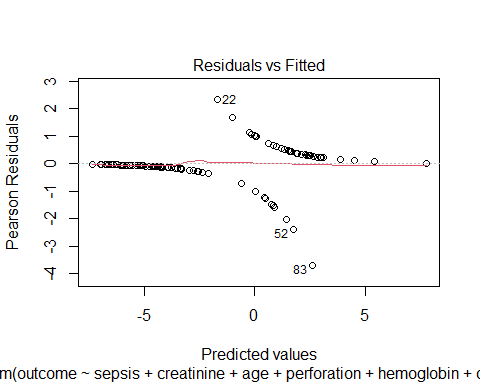
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction alive dead  
## alive 73 10  
## dead 5 33  
##   
## Accuracy : 0.876   
## 95% CI : (0.8038, 0.9289)  
## No Information Rate : 0.6446   
## P-Value [Acc > NIR] : 8.004e-09   
##   
## Kappa : 0.7222   
##   
## Mcnemar's Test P-Value : 0.3017   
##   
## Sensitivity : 0.9359   
## Specificity : 0.7674   
## Pos Pred Value : 0.8795   
## Neg Pred Value : 0.8684   
## Prevalence : 0.6446   
## Detection Rate : 0.6033   
## Detection Prevalence : 0.6860   
## Balanced Accuracy : 0.8517   
##   
## 'Positive' Class : alive   
##

The accuracy is 87.6%

Sensitivity is 93.5

Specificity is 76.7

plot(Model\_Outcome\_ia)



### AUROC

roc <- roc(Data1$outcome,Model\_Outcome\_ia$fitted.values)

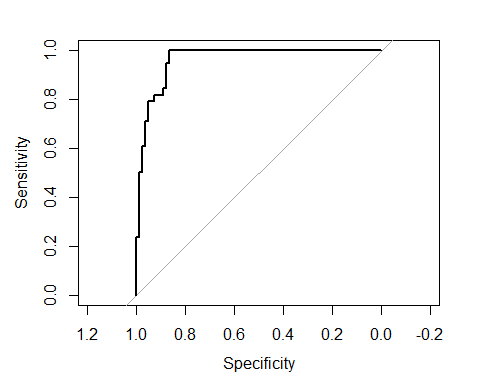
## Setting levels: control = alive, case = dead

## Setting direction: controls < cases

auc(roc)

## Area under the curve: 0.9623

plot(roc)



### Hosmer-Lemeshow

#fit.hl <- gof(Model\_Outcome\_ia, g=8)

#fit.hl$gof

p-value for HL and Oseo Rojek are all > 0.05, hence the model is fit

hltest(Model\_Outcome\_ia, G=10)

##   
## Modified Hosmer-Lemeshow test for large samples  
## H0: epsilon<=epsilon0 vs. Ha: epsilon>epsilon0  
## In-sample goodness of fit.  
##   
## data: Model evaluated on 121 observations  
## HL statistic = 1.8246, dof = 8.00000000, lambda = 0.00090838, p-value =  
## 0.9859  
## alternative hypothesis: true epsilon is greater than 0.002739948  
## 95 percent confidence interval:  
## 0 Inf  
## sample estimates:  
## epsilonHat   
## 0