Regression II Final Project Group 1M

Yingzhi Ma, Guanyu Lu, Yi Yang

2023-11-26

Package may use in the project

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

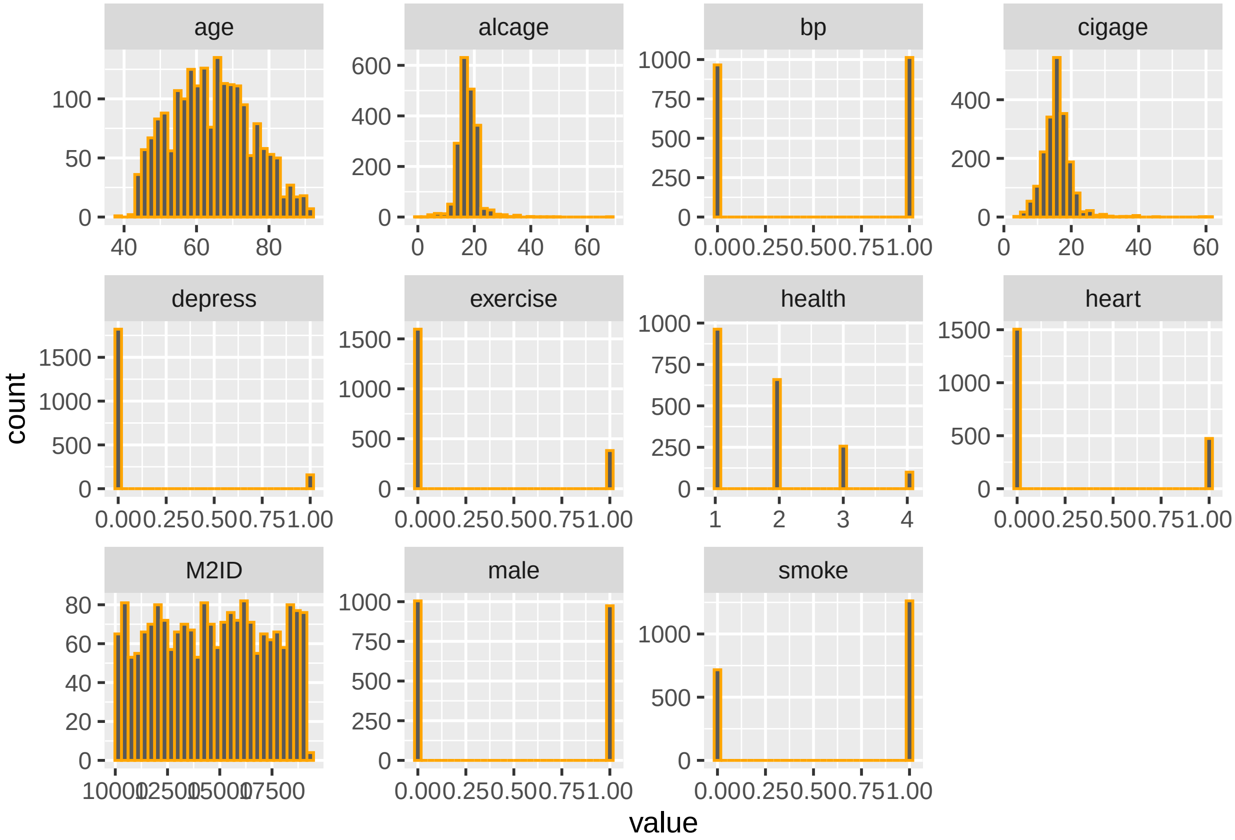
Upload Data from Excel File

```r
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

```r
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‘.
```



```r
#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

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).

```r
#Step 1
my_data_1 <- my_data %>% dplyr::select(heart, bp, smoke, age, male, exercise);
model_1 <- glm(heart ~ bp + smoke + age + male + exercise,
               family = binomial, data = my_data_1);
summary(model_1);
```

```
##
## Call:
## glm(formula = heart ~ bp + smoke + age + male + exercise, family = binomial,
##     data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4202  -0.7818  -0.5533  -0.3407   2.4150
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.909779   0.370817 -13.240  < 2e-16 ***
## bp           0.851829   0.116975   7.282 3.29e-13 ***
## smoke        0.282466   0.119265   2.368   0.0179 *
## age          0.042696   0.005266   8.107 5.18e-16 ***
## male         0.444804   0.111906   3.975 7.04e-05 ***
## exercise     0.189172   0.143432   1.319   0.1872
## ---
## 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: 1993.1  on 1973  degrees of freedom
## AIC: 2005.1
##
## Number of Fisher Scoring iterations: 4
```

```r
#Step 2
model_1_2 <- glm(heart ~ bp + smoke + age + male + exercise + smoke * male,
                 family = binomial, data = my_data_1);
summary(model_1_2);
```

```
##
## Call:
## glm(formula = heart ~ bp + smoke + age + male + exercise + smoke *
##     male, family = binomial, data = my_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4447  -0.7672  -0.5453  -0.3513   2.3915
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.634221   0.384024 -12.068  < 2e-16 ***
## bp           0.849921   0.117204   7.252 4.12e-13 ***
## smoke       -0.028744   0.171281  -0.168   0.8667
## age          0.041602   0.005281   7.878 3.33e-15 ***
## male         0.044772   0.196567   0.228   0.8198
## exercise     0.198499   0.143679   1.382   0.1671
## smoke:male   0.591985   0.239503   2.472   0.0134 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2179  on 1978  degrees of freedom
## Residual deviance: 1987  on 1972  degrees of freedom
## AIC: 2001
##
## 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 + exercise
## Model 2: heart ~ bp + smoke + age + male + exercise + smoke * male
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -996.57
## 2   7 -993.51  1 6.1138    0.01341 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Step 3
exp(cbind(OR = coef(model_1), confint(model_1)));
```

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

```
##                      OR       2.5 %     97.5 %
## (Intercept) 0.007374115 0.003528652 0.01510836
## bp          2.343930900 1.866682664 2.95344508
## smoke       1.326396823 1.051535648 1.67875522
## age         1.043620876 1.032970523 1.05452844
## male        1.560184303 1.253640878 1.94432677
## exercise    1.208248216 0.909222492 1.59626720
```

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

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

```
##                      OR       2.5 %    97.5 %
## (Intercept) 0.009713669 0.004527671 0.0204165
## bp          2.339463083 1.862272131 2.9491119
## smoke       0.971665303 0.696111687 1.3633318
## age         1.042479154 1.031810631 1.0534038
## male        1.045789004 0.710778057 1.5375659
## exercise    1.219571084 0.917320958 1.6120459
## smoke:male  1.807572819 1.130644172 2.8930527
```

```r
anova(model_1,model_1_2);
```

```
## Analysis of Deviance Table
##
## Model 1: heart ~ bp + smoke + age + male + exercise
## Model 2: heart ~ bp + smoke + age + male + exercise + smoke * male
##   Resid. Df Resid. Dev Df Deviance
## 1      1973     1993.1
## 2      1972     1987.0  1   6.1138
```

```r
hoslem_1 <- hoslem.test(model_1$y, model_1_2$fitted.values, g=5);
hoslem_1;
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_1$y, model_1_2$fitted.values
## X-squared = 3.4468, df = 3, p-value = 0.3277
```

```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)));
```

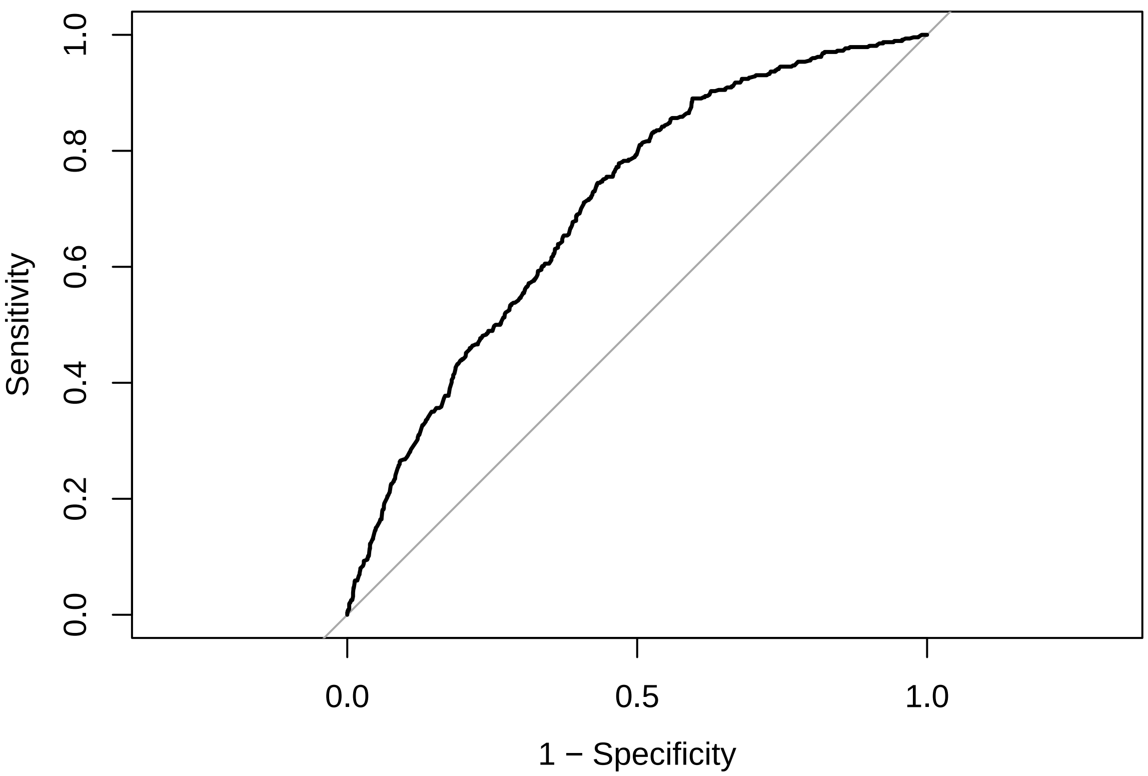
Model_1 AIC	Model_1_2 AIC	Model_1 BIC	Model_1_2 BIC
2005	2001	2039	2040

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

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

## Setting direction: controls < cases
```

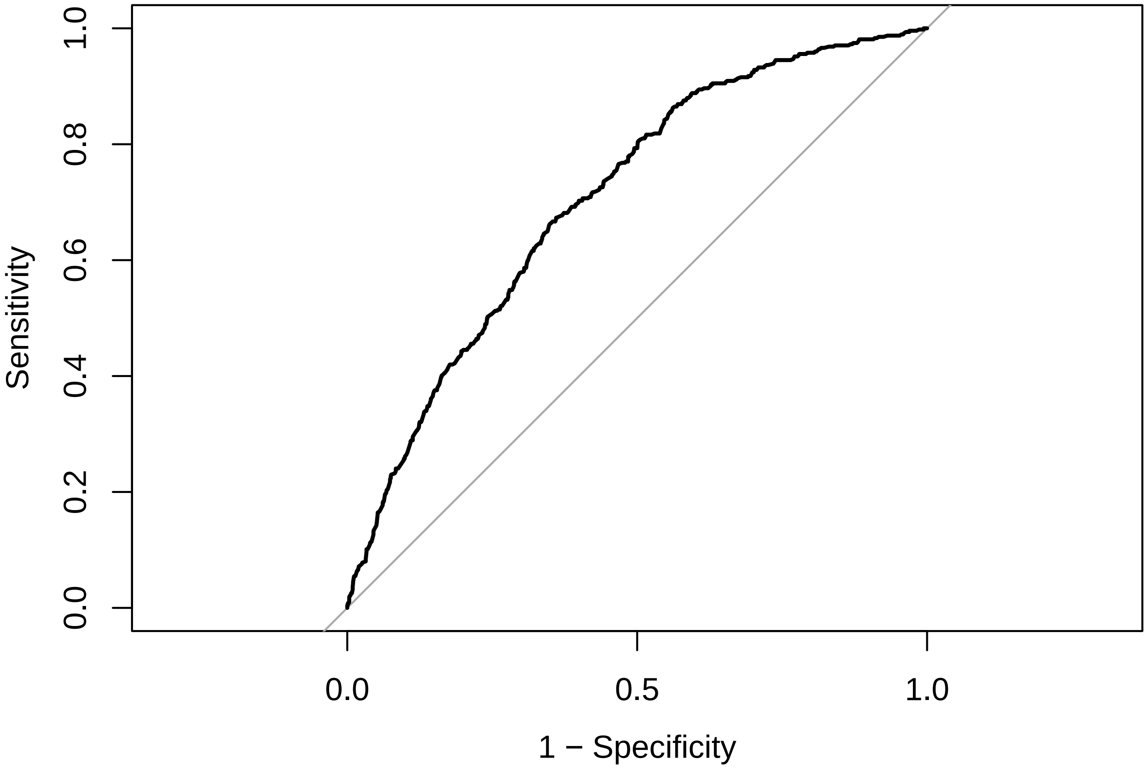
```r
plot.roc(rocplot1, legacy.axes=TRUE);
```



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

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

```r
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
0.7022	0.7062

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) ~ depress + alcage + cigage + age + bp,
                family = cumulative(parallel=TRUE), data = my_data_2);
```

```
## Warning in eval(slot(family, "initialize")): response should be ordinal---see
## ordered()
```

```r
summary(model_2);
```

```
##
## Call:
## vglm(formula = factor(health) ~ depress + alcage + cigage + age +
##     bp, family = cumulative(parallel = TRUE), data = my_data_2)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.870987   0.305714   2.849  0.00439 **
## (Intercept):2  2.547893   0.310849   8.197 2.47e-16 ***
## (Intercept):3  4.033825   0.323450  12.471  < 2e-16 ***
## depress       -1.201805   0.154374  -7.785 6.97e-15 ***
## alcage         0.003656   0.011935   0.306  0.75934
## cigage        -0.005030   0.010421  -0.483  0.62931
## age           -0.005494   0.004165  -1.319  0.18721
## bp            -0.945530   0.090151 -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
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##   depress    alcage    cigage       age        bp
## 0.3006510 1.0036629 0.9949824 0.9945215 0.3884738
```

```r
#Step 2
#Relative Odd Ratio
fit_rr<-exp(coefficients(model_2));
fit_rr;
```

```
## (Intercept):1 (Intercept):2 (Intercept):3       depress        alcage
##     2.3892684    12.7801470    56.4765448     0.3006510     1.0036629
##       cigage          age           bp
##    0.9949824    0.9945215    0.3884738
```

```
#Step 3
my_data_2_3 <- my_data_2 %>% filter(age == 64) %>% filter(bp == 1);
model_2_3 <- vglm(factor(health) ~ depress, family = cumulative(parallel=TRUE), data = my_data_2_3);
```

```
## Warning in eval(slot(family, "initialize")): response should be ordinal---see
## ordered()
```

```
summary(model_2_3);
```

```
##
## Call:
## vglm(formula = factor(health) ~ depress, family = cumulative(parallel = TRUE),
##     data = my_data_2_3)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.2140     0.3784  -0.566 0.571710
## (Intercept):2   1.6309     0.5013   3.253 0.001140 **
## (Intercept):3   2.6449     0.7418   3.565 0.000363 ***
## depress         0.9473     1.1152   0.849 0.395604
## ---
## 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: 70.9572 on 92 degrees of freedom
##
## Log-likelihood: -35.4786 on 92 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## ’(Intercept):3’
##
##
## Exponentiated coefficients:
##  depress
## 2.578798
```

```
result <- predict(model_2_3, my_data_2_3, type="response");

not_equal_to_1_indices <- c()
for (i in nrow(result)) {
  if (sum(result[i,]) != 1) {
    not_equal_to_1_indices <- c(not_equal_to_1_indices, i)
  }
}
not_equal_to_1_indices
```

```
## NULL
```