STAT 471/571/701 Modern Data Mining Midterm

Hongqian Qin

6:00-8:00 pm, Tuesday, Nov. 5th, 2019

Table of Contents

knitr::opts\_chunk$set(echo = TRUE, fig.width = 7, fig.height = 4)  
if(!require("pacman")) install.packages("pacman")

## Loading required package: pacman

pacman::p\_load(dplyr, ggplot2, magrittr, gridExtra, reshape, rmarkdown, leaps, glmnet, knitr, pROC, reshape2, car) #add your packages here

**Name your submission using the scheme:**

LastName\_FirstName.pdf etc.

**For example:** Zhao\_Linda **.rmd**, **.pdf**, **.html** or **.docx**.

Instruction: This exam requires you to use R. It is completely open book/notes/internet. Write your answers using .rmd format and knitr it into one of the html/pdf/docx format. Show your codes, plots or R-output when needed. You can use echo = TRUE to show your codes. If you have trouble formatting the plots, don’t worry about it. We are not looking for pretty solutions, rather to see if you are able to make sense out of data using R. Make sure the compiled pdf/html/docx shows your answers completely and that they are not cut-off. Throughout the exam, you do not need to use any LaTeX or mathematical equations.

**All the answers should be clearly supported by relevant R code.**

Data for Midterm: The data for midterm can be found at:

/canvas/Files/Midterm/AFR\_2012.csv,

/canvas/Files/Midterm/train\_fram.csv, and

/canvas/Files/Midterm/test\_fram.csv.

Midterm Question File can be found at:

/canvas/Files/Midterm/Miderm11\_05\_2019.Rmd.

**Help:** As always skip any part you have trouble with and you may come back to finish it if you have time. Ask one of us for help if you are stuck somewhere for technical issues.

**Electronic Submission:** In the Assignments section, go to the Midterm assignment and upload your completed files: your .rmd file and a compiled file (either a pdf/html/docx).

You can upload multiple files. The folder will be closed at **08:10PM**.

If you have trouble to upload your files, email them to lzhao@wharton.upenn.edu and arunku@wharton.upenn.edu.

**Whenever we ask for test at some level, assume all the model assumptions are satisfied.**

# The adolescent fertility rate (AFR)

The adolescent fertility rate (AFR) is defined as the number of births per 1,000 women of age 15 to 19. While world’s AFR has been decreasing steadily over the years, some countries still have high AFR. Having children this early in life exposes adolescent women to unnecessary risks. Their chance of dying is twice as high as that of women who wait until their 20s to begin childbearing. In addition, early childbearing greatly reduces the likelihood of a girl advancing her education and limits her opportunities for training and employment.

Based on a data set from the Data Bank of the World Bank (<https://databank.worldbank.org/data/home.aspx>), AFR together with other information of 2012 is available. Our goal is to identify important factors associated with AFR. Hope we could give some recommendations to lower the AFR for policymakers.

The data set is AFR\_2012.csv.

|  |  |
| --- | --- |
| Variable | Description |
| mortality.rate | Mortality rate, under-5 (per 1,000 live births) |
| Country | Country name |
| AFR | Adolescent fertility rate (births per 1,000 women ages 15-19) |
| agri.forestry.fish.gdp.pct | Agriculture, forestry, and fishing, value added (% of GDP) |
| industry.gdp.pct | Industry (including construction), value added (% of GDP) |
| CO2 | CO2 emissions (metric tons per capita) |
| fertility.rate | Fertility rate, total (births per woman) |
| GDP | GDP (current USD) |
| GDP.per.capita | GDP per capita (current US$) |
| gdp.grwoth.rate | GDP growth (annual %) |
| gni | GNI, PPP (current international dollar) |
| inflation | Inflation, GDP deflator (annual %) |
| LE | Life expectancy at birth, total (years) |
| population.growth | Population growth (annual %) |
| population | Population, total |
| unemployment | Unemployment, total (% of total labor force)) |
| Continent | Continent |
| Urban.pop | Percentage of urban population |
| Household.consump | Household consumption expenditure in million |
| Forest.area | Percentage of forest |
| Water | Access to improved water source in percentage |
| Food.prod.index | Food production index |
| Arable.land | Arable land per capita |
| Health.expend | Health expenditure percentage of GDP |
| Immunization | DPT Immunization percentage of children |
| Sanitation.faci | Access to improved sanitation facilities in percentage |
| Immunization.measles | Measles Immunization percentage of children |
| Health.exp.pocket | Percentage of out of pocket health expenditure to total health |
| Fixed.tel | Fixed telephone subscriptions per 100 people |
| Mobile.cel | Mobile cellular subscriptions per 100 people |
| Internet.users | Internet users per 100 people |

## Part 1. EDA

### 1) Reading data

Load AFR\_2012.csv. Notice AFR is Adolescent Fertility Rate.

# you need to put the dataset in the same folder  
# where this .rmd file sits.  
data1 <- read.csv("AFR\_2012.csv")  
data1$X <- NULL

**Use data1 from now.**

**i)** How many countries are there in this data?

dim(data1)

## [1] 114 30

There are 114 countries in this data

**ii)** Are there any missing values? If so, remove them. (You can use the function na.omit().)

sum(is.na(data1))

## [1] 0

There is no missing values.

### 2) Summaries

**i)** Which country has the highest AFR and which one has the lowest AFR?

data1$Country[which.max(data1$AFR)]

## [1] Niger  
## 114 Levels: Algeria Argentina Armenia Austria Azerbaijan ... Vietnam

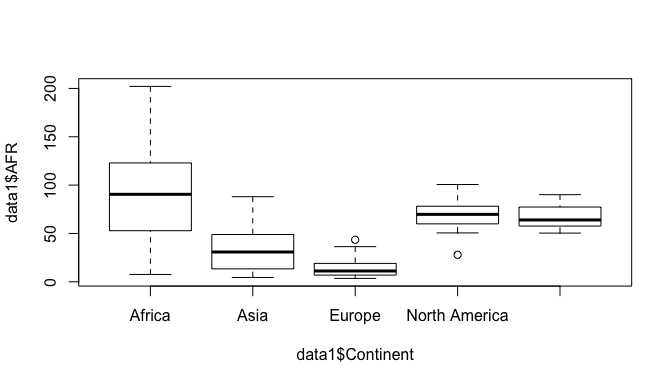
data1$Country[which.min(data1$AFR)]

## [1] Switzerland  
## 114 Levels: Algeria Argentina Armenia Austria Azerbaijan ... Vietnam

Niger has the highest ‘AFR’. Switzerland has the lowest ‘AFR’.

**ii)** Provide a boxplot of AFR among Continent. Comment on the relation between AFR and Continent in 3 sentences.

plot(data1$AFR ~ data1$Continent)

 We find Europe has the lowest median level of AFR and its variance is the smallest, Africa has the highest median level of AFR and its variance is the largest. North America has similar level of AFR compared with South America, while they all significantly higher than the level in Asia.

## Part 2. Analysis with domain knowledge

### 3) AFR vs. a single variable

**i)** Fit a linear model of AFR vs. GDP.per.capita. Is GDP.per.capita significant at 0.01 level? Is the association appearing to be negative?

fit1 <- lm(AFR~GDP.per.capita, data1)  
summary(fit1)

##   
## Call:  
## lm(formula = AFR ~ GDP.per.capita, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.696 -28.628 -7.529 18.814 136.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.8883149 4.3024002 15.314 < 2e-16 \*\*\*  
## GDP.per.capita -0.0011122 0.0001731 -6.425 3.31e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.45 on 112 degrees of freedom  
## Multiple R-squared: 0.2693, Adjusted R-squared: 0.2628   
## F-statistic: 41.28 on 1 and 112 DF, p-value: 3.31e-09

Yes, GDP.per.capita is significant at 0.01 level. The association appearing is negative.

**ii)** Are the averages of AFR the same across all the Continents at 0.01 level? Which continent has the highest AFR on average?

fit2 <- lm(AFR ~ Continent, data1)  
summary(fit2)

##   
## Call:  
## lm(formula = AFR ~ Continent, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -85.746 -9.475 -2.582 12.226 108.773   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 93.336 5.364 17.402 < 2e-16 \*\*\*  
## ContinentAsia -59.795 7.786 -7.680 7.33e-12 \*\*\*  
## ContinentEurope -79.334 7.365 -10.771 < 2e-16 \*\*\*  
## ContinentNorth America -25.032 10.480 -2.388 0.0186 \*   
## ContinentSouth America -25.314 11.307 -2.239 0.0272 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.86 on 109 degrees of freedom  
## Multiple R-squared: 0.5478, Adjusted R-squared: 0.5312   
## F-statistic: 33.01 on 4 and 109 DF, p-value: < 2.2e-16

Anova(fit2)

## Anova Table (Type II tests)  
##   
## Response: AFR  
## Sum Sq Df F value Pr(>F)   
## Continent 117764 4 33.014 < 2.2e-16 \*\*\*  
## Residuals 97204 109   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

No. From the Anova table, the p-value is 2.2e-16. Hence, we have strong evidence to reject the null hypotheses of the averages of AFR are the same across all the Continents at 0.01 level Africa has the highest AFR on average.

### 4) AFR vs GDP.per.capita and Continent

**i)** Fit a linear model of AFR vs GDP.per.capita and Continent, assuming there is no interaction effect.

*a)* Is GDP.per.capita significant at 0.01 level controlling for Continent?

fit3 <- lm(AFR ~ GDP.per.capita + Continent, data1)  
summary(fit3)

##   
## Call:  
## lm(formula = AFR ~ GDP.per.capita + Continent, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -84.850 -10.116 -1.985 11.456 107.944   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.435e+01 5.222e+00 18.066 < 2e-16 \*\*\*  
## GDP.per.capita -4.607e-04 1.677e-04 -2.746 0.00706 \*\*   
## ContinentAsia -5.650e+01 7.657e+00 -7.379 3.47e-11 \*\*\*  
## ContinentEurope -6.553e+01 8.742e+00 -7.497 1.93e-11 \*\*\*  
## ContinentNorth America -2.130e+01 1.027e+01 -2.074 0.04046 \*   
## ContinentSouth America -2.260e+01 1.103e+01 -2.049 0.04286 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29 on 108 degrees of freedom  
## Multiple R-squared: 0.5773, Adjusted R-squared: 0.5578   
## F-statistic: 29.5 on 5 and 108 DF, p-value: < 2.2e-16

GDP.per.capita is significant at 0.01 level controlling for Continent because its p-value is 0.00706.

*b)* Is Continent significant at 0.01 level controlling for GDP.per.capita. For a given GDP.per.capita value, which continent seems to have the lowest AFR on average?

Anova(fit3)

## Anova Table (Type II tests)  
##   
## Response: AFR  
## Sum Sq Df F value Pr(>F)   
## GDP.per.capita 6345 1 7.5422 0.007062 \*\*   
## Continent 66212 4 19.6759 3.456e-12 \*\*\*  
## Residuals 90859 108   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Continent is significant at 0.01 level controlling for GDP.per.capita. For a given GDP.per.capita value, Europe seems to have the lowest AFR on average

**ii)** Some summary statistics seem to indicate a possible interaction effect of Continent and GDP.per.capita over AFR. Run a linear model of AFR vs GDP.per.capita and Continent with interaction.

*a)* Can we reject the null hypothesis of no interaction effect at 0.01 level?

fit4 <- lm(AFR ~ GDP.per.capita \* Continent, data1)  
summary(fit4)

##   
## Call:  
## lm(formula = AFR ~ GDP.per.capita \* Continent, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -87.447 -10.153 -1.849 8.412 86.525   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 120.419266 6.048705 19.908  
## GDP.per.capita -0.012350 0.001882 -6.563  
## ContinentAsia -76.166753 8.513451 -8.947  
## ContinentEurope -99.189984 8.951556 -11.081  
## ContinentNorth America -42.199062 11.167372 -3.779  
## ContinentSouth America -45.140165 18.582881 -2.429  
## GDP.per.capita:ContinentAsia 0.011204 0.001925 5.822  
## GDP.per.capita:ContinentEurope 0.012125 0.001888 6.421  
## GDP.per.capita:ContinentNorth America 0.011387 0.001963 5.802  
## GDP.per.capita:ContinentSouth America 0.011453 0.002689 4.260  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## GDP.per.capita 2.11e-09 \*\*\*  
## ContinentAsia 1.51e-14 \*\*\*  
## ContinentEurope < 2e-16 \*\*\*  
## ContinentNorth America 0.000263 \*\*\*  
## ContinentSouth America 0.016850 \*   
## GDP.per.capita:ContinentAsia 6.53e-08 \*\*\*  
## GDP.per.capita:ContinentEurope 4.12e-09 \*\*\*  
## GDP.per.capita:ContinentNorth America 7.15e-08 \*\*\*  
## GDP.per.capita:ContinentSouth America 4.50e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.62 on 104 degrees of freedom  
## Multiple R-squared: 0.7067, Adjusted R-squared: 0.6813   
## F-statistic: 27.84 on 9 and 104 DF, p-value: < 2.2e-16

We can reject the null hypothesis of no interaction effect at 0.01 level

## Part 3. Analysis with LASSO

Lastly we will build a parsimonious model to see what factors are related to AFR.

### 5) LASSO to reduce the number of factors

**i)** In any linear model you will run, can you include the variable Country in it? Why or Why not? Explain in no more than 2 sentences. No. Because each country is a variable, you don’t really get meaningful results from 114 levels

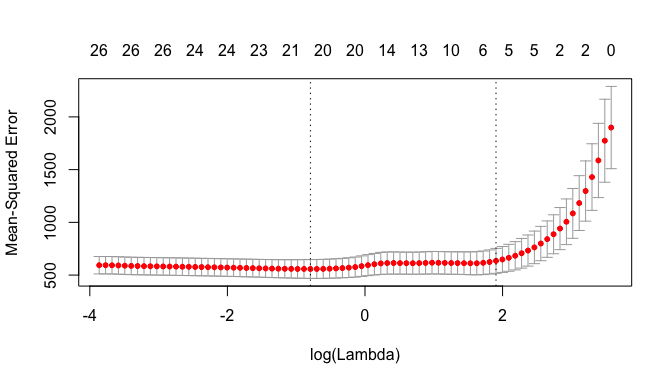
We now take out Country, fertility.rate, Continent and save it as data2.

data2 <- data1 %>% dplyr::select(-Country, -fertility.rate, -Continent)

**ii)** LASSO with cv.glmnet

*a)* Run a LASSO analysis using all variables in data2. For reproducibility, use set.seed(1). Also use 10 folds by setting nfolds=10. Plot the LASSO output.

set.seed(1)  
X <- model.matrix(AFR~.,data=data2)[,-1]   
Y <- data1$AFR  
fit.lasso.0 <- cv.glmnet(X,Y, nfolds = 10)   
plot(fit.lasso.0)

 *b)* Choose 6 non-zero variables from LASSO. **Hint:** The top line in the plot shows the number of non-zero coefficients. Choose approximately equal to exponential of value on x-axis that corresponds to in the top line.

coef.nzero <- coef(fit.lasso.0, nzero = 6)  
coef.nzero <- coef.nzero[which(coef.nzero !=0), ]  
rownames(as.matrix(coef.nzero))

## [1] "(Intercept)" "mortality.rate" "CO2" "Water"   
## [5] "Immunization" "Sanitation.faci" "Internet.users"

### 6) Final analysis using variables from LASSO

**i)** Assume we obtain the following variables from LASSO: mortality.rate, Water, Immunization, Sanitation.faci. Run the final linear model of AFR with the variables listed here AND Continent. Call this fit fit\_final\_AFR. Report the Anova of this fit and report if any of the variables are insignificant at 0.05 level.

**Note: data2 does not contain Continent. Also, we are giving the variables so that students who are not able to output LASSO variables will not be double penalized. This may not be the variables from the LASSO output.**

fit\_final\_AFR <-lm(AFR ~ mortality.rate+Water+Immunization+Sanitation.faci+Continent, data=data1)  
summary(fit\_final\_AFR)

##   
## Call:  
## lm(formula = AFR ~ mortality.rate + Water + Immunization + Sanitation.faci +   
## Continent, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.090 -6.926 -1.108 8.960 65.139   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 164.2197 33.0370 4.971 2.61e-06 \*\*\*  
## mortality.rate 0.4835 0.1354 3.571 0.000539 \*\*\*  
## Water -0.4954 0.2326 -2.130 0.035524 \*   
## Immunization -0.6875 0.2577 -2.668 0.008844 \*\*   
## Sanitation.faci -0.3702 0.1438 -2.574 0.011443 \*   
## ContinentAsia -4.9808 6.4118 -0.777 0.439014   
## ContinentEurope -3.2148 7.7039 -0.417 0.677319   
## ContinentNorth America 36.3918 8.1479 4.466 2.01e-05 \*\*\*  
## ContinentSouth America 31.6557 8.5002 3.724 0.000317 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 105 degrees of freedom  
## Multiple R-squared: 0.8371, Adjusted R-squared: 0.8247   
## F-statistic: 67.43 on 8 and 105 DF, p-value: < 2.2e-16

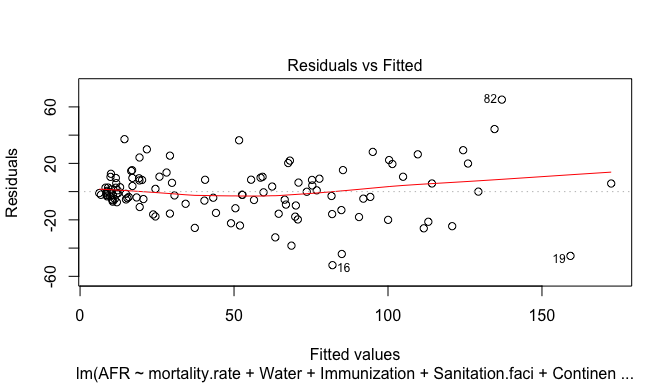
Anova(fit\_final\_AFR)

## Anova Table (Type II tests)  
##   
## Response: AFR  
## Sum Sq Df F value Pr(>F)   
## mortality.rate 4253 1 12.7503 0.0005386 \*\*\*  
## Water 1513 1 4.5361 0.0355243 \*   
## Immunization 2374 1 7.1175 0.0088437 \*\*   
## Sanitation.faci 2210 1 6.6263 0.0114428 \*   
## Continent 22633 4 16.9636 9.308e-11 \*\*\*  
## Residuals 35023 105   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

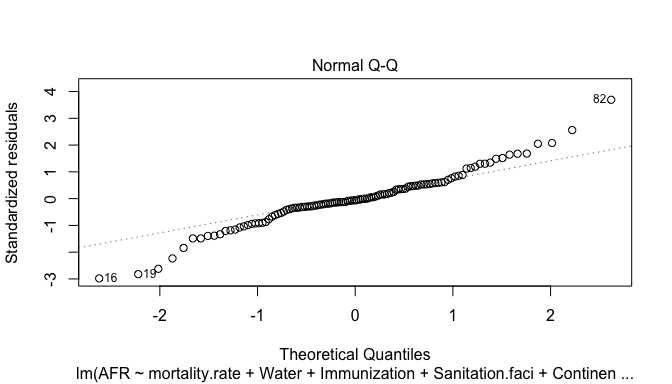
ContinentAsia and ContinentEurope are not significant at 0.05 level. While Anova result shows all variables are significant.

**ii)** By your judgement, are the linear model assumptions satisfied for fit\_final\_AFR? Provide relevant plots.

plot(fit\_final\_AFR, 1)



plot(fit\_final\_AFR, 2)

 The linear model assumptions are not satisfied.

**iii)** Based on the summary of fit\_final\_AFR, provide one variable in this which the policy makers can use to decrease AFR. 1. Improve access to safe water resources 2. Increase DPT immunization percentage of children 3. Increase access to improved sanitation facilities in percentage 4. Try to decrease the mortality.rate

$$
\mbox{\textbf{End of AFR analysis}}.
$$

# Relation between Heart Disease and Smoking

In this part, we will explore the relation between heart disease and smoking using Framingham dataset. As we saw in class this dataset has a factor variable of interest HD which takes values “0” or “1” with “1” indicating the presence of heart disease. It includes other variables such as AGE, SEX, SBP, DBP, CHOL, FRW and CIG. We will use a revised data for the purpose of the midterm. A new categorical variable Smoke is created by grouping the original continuous variable CIG into categories “None”, “Med”, “High” and “VHigh”. We have split the original Framingham dataset into training and testing data: train\_fram.csv and test\_fram.csv.

## load the dataset train\_fram.csv and testing data here   
HD\_train <- read.csv("train\_fram.csv")  
HD\_train$HD <- factor(HD\_train$HD, levels = c("0", "1"))  
HD\_train$Smoke <- factor(HD\_train$Smoke, levels = c("None", "Med", "High", "VHigh"))  
HD\_train$X <- NULL

## Part 1 Relation between HD and Smoke

### 1) Preliminary Models

**i)** Fit a logistic regression between HD and Smoke. Call this model fit1\_logi. Report the summary. What is the base level? At what level/category of Smoke, the probability of HD = 1 appears to be the highest?

fit1\_logi <- glm(HD~Smoke, HD\_train, family=binomial(logit))  
summary(fit1\_logi, results=TRUE)

##   
## Call:  
## glm(formula = HD ~ Smoke, family = binomial(logit), data = HD\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8702 -0.6740 -0.6740 -0.5945 1.9081   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.3665 0.1045 -13.081 <2e-16 \*\*\*  
## SmokeMed -0.2771 0.2506 -1.106 0.2688   
## SmokeHigh 0.3464 0.1913 1.811 0.0701 .   
## SmokeVHigh 0.5907 0.2475 2.386 0.0170 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1053.8 on 999 degrees of freedom  
## Residual deviance: 1043.2 on 996 degrees of freedom  
## AIC: 1051.2  
##   
## Number of Fisher Scoring iterations: 4

The base level is smoke. SmokeVHigh has the highest probability of causing ‘HD=1’

**ii)** In model fit1\_logi, is Smoke a significant variable at level 0.05?

Anova(fit1\_logi)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: HD  
## LR Chisq Df Pr(>Chisq)   
## Smoke 10.655 3 0.01375 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Smoke is a significant variable at level 0.05

**iii)** Now fit a logistic regression model for HD using AGE, SEX, SBP, CHOL and Smoke as covariates/features. Let us call this model fit2\_logi. Is Smoke a significant variable at level 0.05?

fit2\_logi <- glm(HD~AGE+SEX+SBP+CHOL+Smoke, HD\_train, family=binomial(logit))  
summary(fit2\_logi, results=TRUE)

##   
## Call:  
## glm(formula = HD ~ AGE + SEX + SBP + CHOL + Smoke, family = binomial(logit),   
## data = HD\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5165 -0.7313 -0.5566 -0.3549 2.5273   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.875653 1.080740 -7.287 3.16e-13 \*\*\*  
## AGE 0.052295 0.017208 3.039 0.00237 \*\*   
## SEXMALE 0.909009 0.183307 4.959 7.09e-07 \*\*\*  
## SBP 0.016669 0.002852 5.845 5.06e-09 \*\*\*  
## CHOL 0.003286 0.001770 1.857 0.06338 .   
## SmokeMed -0.275544 0.261996 -1.052 0.29293   
## SmokeHigh 0.224530 0.211482 1.062 0.28837   
## SmokeVHigh 0.438044 0.275931 1.588 0.11240   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1053.82 on 999 degrees of freedom  
## Residual deviance: 968.11 on 992 degrees of freedom  
## AIC: 984.11  
##   
## Number of Fisher Scoring iterations: 4

Anova(fit2\_logi)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: HD  
## LR Chisq Df Pr(>Chisq)   
## AGE 9.349 1 0.00223 \*\*   
## SEX 25.309 1 4.885e-07 \*\*\*  
## SBP 34.999 1 3.299e-09 \*\*\*  
## CHOL 3.430 1 0.06401 .   
## Smoke 5.380 3 0.14599   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Smoke is NOT a significant variable at level 0.05

## Part 2: Classification

### 2) Thresholding Rules

Load the testing data test\_fram.csv.

HD\_test <- read.csv("test\_fram.csv")  
HD\_test$HD <- factor(HD\_test$HD, levels = c("0", "1"))  
HD\_test$Smoke <- factor(HD\_test$Smoke, levels = c("None", "Med", "High", "VHigh"))  
HD\_test$X <- NULL

**i)** Use the 1/2 thresholding rule for predicting HD with models fit1\_logi and fit2\_logi. Predict HD on the testing data. What are the (testing) misclassification errors from models fit1\_logi and fit2\_logi? Report at least 3 decimals.

fit1\_logi.pred <- predict(fit1\_logi, HD\_test, type = "response") >= 0.5   
mean(as.numeric(fit1\_logi.pred) != HD\_test$HD)

## [1] 0.221374

fit2\_logi.pred <- predict(fit2\_logi, HD\_test, type = "response") >= 0.5   
mean(as.numeric(fit2\_logi.pred) != HD\_test$HD)

## [1] 0.2188295

**ii)** Based on the testing MCE, which model is the best? fit2\_logi is better because it has smaller MCE.

## Part 3: Prediction

### 3) Prediction

**i)** We have a male with features: AGE = 50, SBP = 160, CHOL = 230 and Smoke = None. Predict whether this person has a heart disease or not based on the 1/2 thresholding rule with fit1\_logi.

fit1\_logi

##   
## Call: glm(formula = HD ~ Smoke, family = binomial(logit), data = HD\_train)  
##   
## Coefficients:  
## (Intercept) SmokeMed SmokeHigh SmokeVHigh   
## -1.3665 -0.2771 0.3464 0.5907   
##   
## Degrees of Freedom: 999 Total (i.e. Null); 996 Residual  
## Null Deviance: 1054   
## Residual Deviance: 1043 AIC: 1051

Because the male does not smoke, according to the fit1\_logi model, he will not have a heart disease.

$$
\mbox{\textbf{End of the exam}}.
$$

# Declaration

By submitting this document you certify that you have complied with the University of Pennsylvania’s Code of Academic Integrity, to the best of your knowledge. You further certify that you have taken this exam under its sanctioned conditions, i.e. solely within the set exam room and within the time allotted.