Midterm

STAT 471/571/701 Modern Data Mining

6:00-8:00 pm, Tuesday, Nov. 5th, 2019

Contents

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. 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.

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.

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.

```r
# 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?

```r
length(unique(data1$Country))
```

```
## [1] 114
```

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

```r
sum(is.na(data1))
```

```
## [1] 0
```

2) Summaries

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

```r
data1[data1$AFR == min(data1$AFR),"Country"]
```

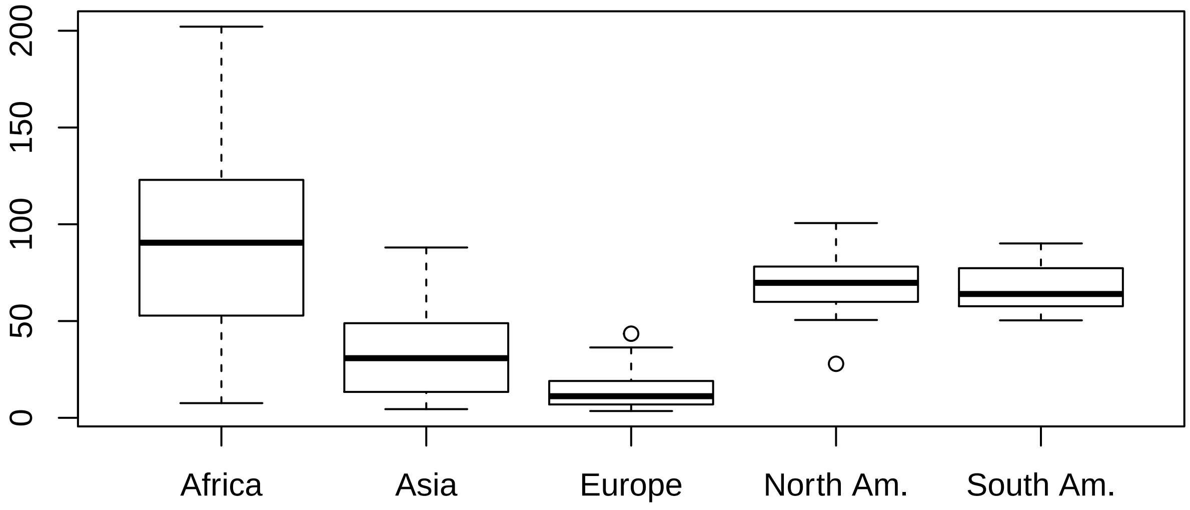
```
## [1] Switzerland
## 114 Levels: Algeria Argentina Armenia Austria Azerbaijan ... Vietnam
```

```r
data1[data1$AFR == max(data1$AFR),"Country"]
```

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

ii) Provide a boxplot of `AFR` among `Continent`. Comment on the relation in one sentence.

```r
boxplot(data1$AFR[data1$Continent == "Africa"], data1$AFR[data1$Continent == "Asia"],
        data1$AFR[data1$Continent == "Europe"],
        data1$AFR[data1$Continent == "North America"],
        data1$AFR[data1$Continent == "South America"],
        names = c("Africa", "Asia", "Europe", "North Am.", "South Am."))
```



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?

```
summary(lm(AFR ~ GDP.per.capita, data = data1))
```

```
##
## 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
```

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

```
Anova(lm(AFR ~ Continent, data = data1))
```

```
## 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
```

```
summary(lm(AFR ~ Continent, data = data1))
```

```
##
## 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
```

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`?

```r
Anova(lm(AFR ~ GDP.per.capita + Continent, data = data1))
```

```
## 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
```

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

```r
summary(lm(AFR ~ GDP.per.capita + Continent, data = data1))
```

```
##
## 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
```

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?

```
Anova(lm(AFR ~ GDP.per.capita*Continent, data = data1))
```

```
## Anova Table (Type II tests)
## 
## Response: AFR
##                          Sum Sq  Df F value    Pr(>F)    
## GDP.per.capita             6345   1  10.465  0.001631 ** 
## Continent                 66212   4  27.302 1.696e-15 ***
## GDP.per.capita:Continent  27804   4  11.465 9.516e-08 ***
## Residuals                 63055 104                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(AFR ~ GDP.per.capita*Continent, data = data1))
```

```
## 
## 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
```

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 `Country` in it? Why or Why not? Explain in no more than 2 sentences. (No points if you write more.)

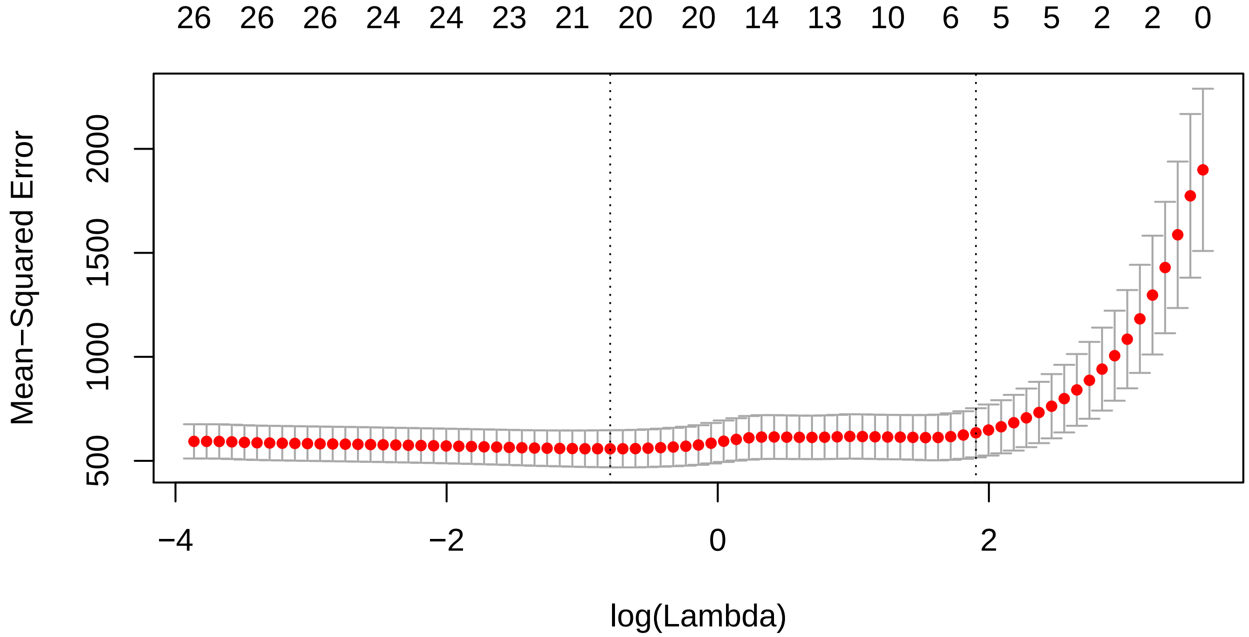
We now take out `Country`, `fertility.rate`, `Continent` and save it as `data2`.

```
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.

```
dat <- model.matrix(AFR ~ ., data = data2)
set.seed(1)
fit_cv <- cv.glmnet(dat[,-1],data2$AFR,nfolds = 10)
plot(fit_cv)
```



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

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`. Report the Anova of this fit and report if any of the variables are insignificant at 0.05 level.

Note: data2 does not 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 true set of the LASSO output.

```r
fit_final <- lm(AFR ~ mortality.rate + Water + Immunization + Sanitation.faci + Continent, data = data1
Anova(fit_final)
```

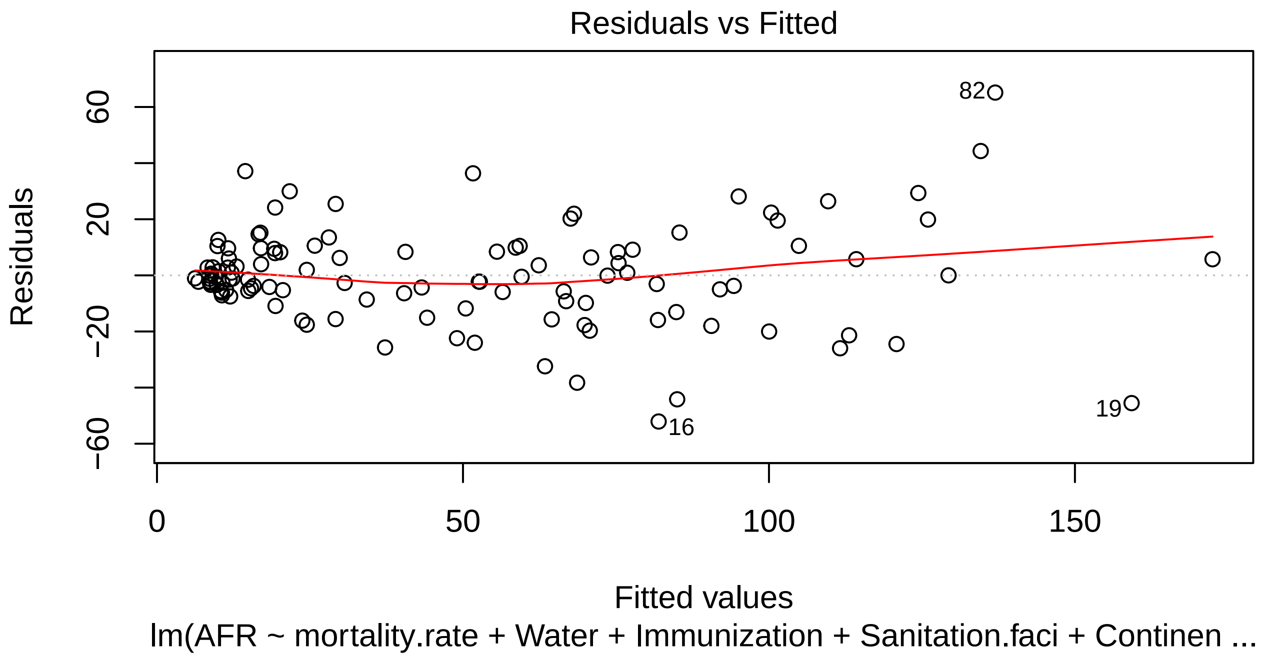
```
## Anova Table (Type II tests)
## 
## Response: AFR
##                 Sum Sq  Df F value    Pr(>F)    
## mortality.rate    4253   1 12.7503 0.0005386 ***
## Water             1513   1  4.5361 0.0355243 *  
## Immunization      2374   1  7.1175 0.0088437 ** 
## Sanitation.faci   2210   1  6.6263 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
```

```r
summary(fit_final)
```

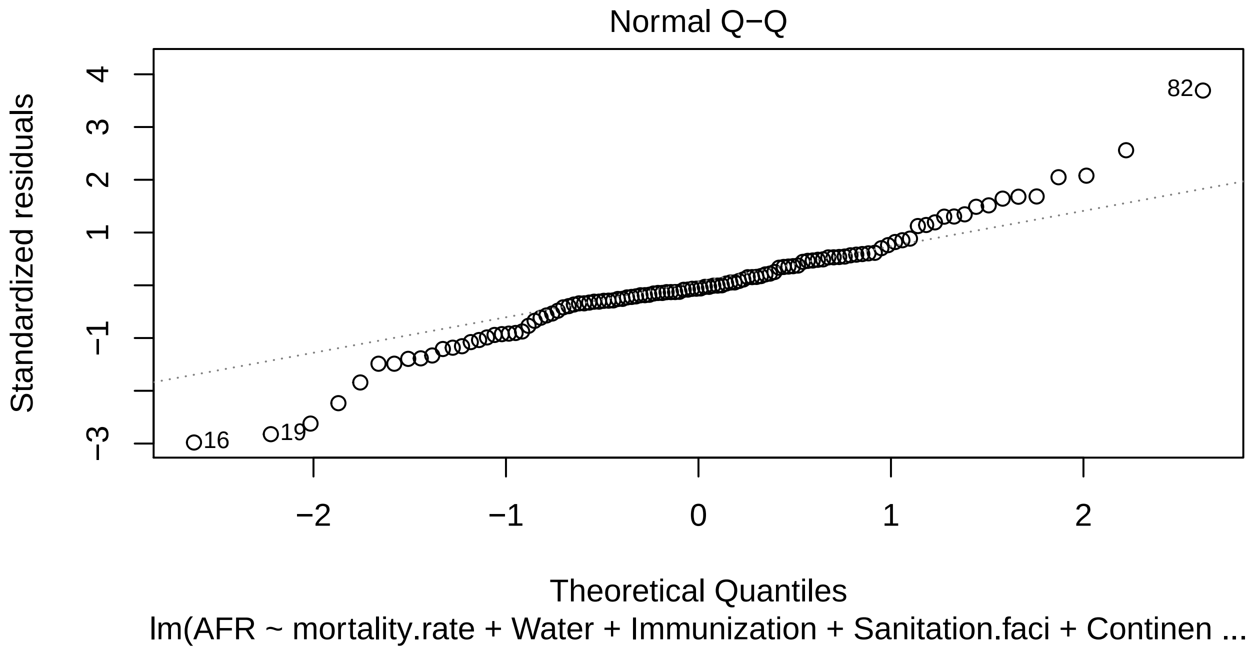
```
## 
## Call:
## lm(formula = AFR ~ mortality.rate + Water + Immunization + Sanitation.faci + 
##     Continent, data = data1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -52.090  -6.926  -1.108   8.960  65.139 
## 
## Coefficients:
##                         Estimate Std. Error 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
```

ii) Use no more than 4 sentences to summarize your findings (including validity of linear model assumptions). No points if you write more than 4 sentences.

```r
plot(fit_final, 1)
```



```r
plot(fit_final, 2)
```



End of PART I.

Relation between Heart Disease and Smoking

In this part, we will explore the relation between heart disease and smoking using Framingham dataset. This is not the same dataset used in class but is revised for the purpose of the midterm. A new categorical variable `Smoke` is created by grouping the orginal continuous varialbe `CIG`. We have split the original Framingham dataset into training and testing data: `HD_train` and `HD_test`.

NOTE:

```
## load the dataset train_fram.csv and testing data here
HD_train <- read.csv("train_fram.csv")
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, data = HD_train, family = binomial)
summary(fit1_logi)
```

```
##
## Call:
## glm(formula = HD ~ Smoke, family = binomial, 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
```

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
```

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, family = binomial, data = HD_train)
# summary(fit2_logi)
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
```

Part 2: Classification

2) Thresholding Rules

i) Load the testing data `test_fram.csv`. 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 three decimals.

```
HD_test <- read.csv("test_fram.csv")
prob_fit1 <- fit1_logi %>% predict(HD_test, type = "response")
pred_fit1 <- ifelse(prob_fit1 > 0.5, 1, 0)
mean(pred_fit1 != HD_test$HD)
```

```
## [1] 0.221374
```

```
prob_fit2 <- fit2_logi %>% predict(HD_test, type = "response")
pred_fit2 <- ifelse(prob_fit2 > 0.5, 1, 0)
mean(pred_fit2 != HD_test$HD)
```

```
## [1] 0.2188295
```

ii) Based on the testing MCE, which model is the best?

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`.

```
newdata <- data.frame(AGE = 50, SBP = 160, CHOL = 230, Smoke = "None", SEX = "MALE")
prob_fit1_predict <- fit1_logi %>% predict(newdata, type = "response")
HD_fit1_predict <- ifelse(prob_fit1_predict > 0.5, 1, 0)
# prob_fit2_predict <- fit2_logi %>% predict(newdata, type = "response")
# HD_fit2_predict <- ifelse(prob_fit2_predict > 0.5, 1, 0)
prob_fit1_predict
```

```
##         1
## 0.2031802
```

```
# prob_fit2_predict
HD_fit1_predict
```

```
## 1
## 0
```

```
# HD_fit2_predict
```

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.