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. 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.

```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?

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

2) Summaries

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

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

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?

ii) Are the averages of `AFR` the same across all the Continents at 0.01 level? Which continent 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`?

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?

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?

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.

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.

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`. 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.

ii) By your judgement, are the linear model assumptions satisfied for `fit_final_AFR`? Provide relevant plots.

iii) Based on the summary of `fit_final_AFR`, provide one variable in this which the policy makers can use to decrease `AFR`.

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`.

```r
## 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?

ii) In model `fit1_logi`, is `Smoke` 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?

Part 2: Classification

2) Thresholding Rules

Load the testing data `test_fram.csv`.

```r
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.

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`.

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.