Midterm

STAT 471/571/701 Modern Data Mining

03/25/2019

Contents

Part I: Mortality rate under age five . . . 3
Question 1: EDA of `data` . . . 4
a) Quick Summary . . . 4
b) `mortality.rate` in 2012 . . . 4
Question 2: Relation between mortality.rate and other variables . . . 4
a) Single most usueful factor . . . 4
b) `mortality.rate` vs. `GDP` . . . 5
c) Relation between mortality.rate and other variables. . . . 5
Question 3: Linear Model building . . . 5
a) LASSO Regression: fit.lasso.0 . . . 5
b) Lasso fit . . . 6
Question 4: Prediction intervals . . . 6
Part II: Logistic regression/classification, Breast Cancer Prediction . . . 7
Question 1: Data preparation . . . 7
Question 2: Linear versus Logistic Regression . . . 7
a) Logistic regression fit.glm . . . 7
b) Linear regreesioin fit.lm . . . 8
c) Findings . . . 8

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. Write your answers using .rmd format and knitr it into one of the html/pdf/docx format. **It looks like the easiest way is to knit it to a .html file!** 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/mortality_2012.csv` and `/canvas/Files/Midterm/breast-cancer.csv`.

Midterm Question File can be found at:

`/canvas/Files/Midterm/Miderm03_25_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`.

Answering the questions: For all the questions whenever asked, you answer should show the R code first and then a description of the results to answer the questions. Do not print numbers without context from R code chunk.

Example: If you asked to report the mean of `mortality.rate`, then your answer should resemble

```
# you need to put the dataset in the same folder where this .rmd file sits.
data <- read.csv("mortality_2012.csv")
cat("Mean of the mortality rate is:\n")
```

```
## Mean of the mortality rate is:
```

```
mean(data$mortality.rate)
```

```
## [1] 31.72783
```

or

```
mean(data$mortality.rate)
```

```
## [1] 31.72783
```

The mean of the mortality rate is 31.72783.

Do not just write

```
mean(data$mortality.rate)
```

```
## [1] 31.72783
```

without any context or sentence answering the question.

Part I: Mortality rate under age five

According to World Health Organization (WHO), 5.4 million children under age five died in 2017. The risk of a child dying before completing five years of age is still highest in Africa, 8 times compared to that in Europe. In addition, gaps of child mortality between high-income and low-income countries remain large. Reducing these inequalities across countries and saving more child lives by ending preventable child deaths are important priorities of WHO.

In this exam, we will look into the mortality rate of children under age five of 115 countries around the world in 2012. The goal is to identify important factors associated with children mortality rate and to be able to quantify the relationship.

The data is obtained from DataBank of the World Bank. https://databank.worldbank.org/data/home.aspx. The following R-chunk reads the data `mortality_2012.csv`.

```
# you need to put the dataset in the same folder where this .rmd file sits.
data <- read.csv("mortality_2012.csv")
```

Variable	Description
mortality.rate	Mortality rate, under-5 (per 1,000 live births)
Country	Country name
adolescent.fertility.rate	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 US$)
GDP.per.capita	GDP per capita (current US$)
gdp.grwoth.rate	GDP growth (annual %)
gni	GNI, PPP (current international $)
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

Question 1: EDA of `data`

a) Quick Summary

Report the following information about `data`:

i) How many variables and observations does `data` have?

ii) Are there any missing values?

b) `mortality.rate` in 2012

i) Which country has the highest `mortality.rate`? And which country has the lowest `mortality.rate`? What are the mean and median `mortality.rate` among all countries in data?

ii) Make a histogram of the `mortality.rate`. Use no more than three sentences to describe the distribution of the `mortality.rate`. (Does it look normal? Are there more countries with low `mortality.rate` or more countries with high `mortality.rate`?)

iii) Report the mean and median `mortality.rate` by `Continent`. Which `Continent` has the highest mean `mortality.rate` and what is the value?

iv) Show the boxplots of `mortality.rate` versus `Continent`. Write a brief summary based on these boxplots. No more than three sentences please.

Question 2: Relation between mortality.rate and other variables

a) Single most usueful factor

```
cor.mat <- data %>% select(-Country, -Continent) %>% cor
cor.mat[1,]
```

```
##             mortality.rate  adolescent.fertility.rate
##                1.000000000               0.826333224
## agri.forestry.fish.gdp.pct          industry.gdp.pct
##                0.811744183              -0.105831654
##                        CO2            fertility.rate
##               -0.578322767               0.905627614
##                        GDP            GDP.per.capita
##               -0.188500283              -0.483895183
##            GDP.grwoth.rate                       GNI
##                0.469894159              -0.118341374
##                  inflation         population.growth
##                0.071930261               0.491870293
##                 population              unemployment
##               -0.006172731              -0.166284326
##                  Urban.pop         Household.consump
##               -0.675863608              -0.507710451
##                Forest.area               Arable.land
##               -0.124441650              -0.011832658
##                       Water           Food.prod.index
##               -0.803717025               0.479603735
##              Health.expend              Immunization
##               -0.490577957              -0.729817153
##            Sanitation.faci       Immunization.measles
```

```
##               -0.887278300               -0.731229148
##          Health.exp.pocket                  Fixed.tel
##                0.406867337               -0.692452102
##                 Mobile.cel             Internet.users
##               -0.641432817               -0.777111982
```

i) What does `cor.mat` contain? And what does `cor.mat[1,]` show?

ii) Based on the above correlation matrix, which single continuous variable will have the highest R^2 if we fit `mortality.rate` vs one variable at a time and why? **We only take the answer based on the above R-chunk! No need to do all the simple linear regressions.**

b) `mortality.rate` vs. `GDP`

i) Fit a linear model of `mortality.rate` vs. GDP. Make a scatter plot of GDP vs. `mortality.rate`, together with the regression line overlayed. Report the lm summary statistics. Is GDP a significant variable at .01 level?

ii) Fit a linear model of `mortality.rate` vs. log(GDP). Here we use natural log. Make another scatter plot of log(GDP) vs. `mortality.rate` together with the regression liner. Report the lm summary statistics. Is the GDP in log scale significant at .01 level?

iii) Which is a better model choice? And why? No more than three sentences.

iv) Use your model in **ii)** regardless your answer in **iii)** and write your findings briefly (no more than 3 lines) summarizing the relationship between GDP and `mortality.rate`.

c) Relation between mortality.rate and other variables.

Now examine the relationship between mortality.rate vs Sanitation.faci, log(GDP) and Continent.

i) Fit a model of `mortality.rate` vs. Log(GDP), Sanitation.faci and Continent. Report the summary.

ii) Is log(GDP) a significant variable at .01 level after controlling for `Sanitation.faci` and `Continent`?

iii) Are the means of mortality.rate among all Continents the same at .05 level after controlling for log(GDP) and `Sanitation.faci`?

iv) Based on this model fit, which Continent appears to have the highest mortality.rate after controlling for log(GDP) and `Sanitation.faci`? (No test needed.)

Question 3: Linear Model building

In this question, we build a model for `mortality.rate` based on the covariates available in data. Your professor insists on that GDP should have been taken a log scale. So from now on you may drop GDP from the working data but keep log(GDP) there to avoid any potential issues. Call this extracted data as `data1`. (Show your code for this.)

a) LASSO Regression: fit.lasso.0

i) Country names should not be a predictor. Explain why not? (one sentence only)

ii) LASSO Regression: Use `cv.glmnet()` function on the data for the response `mortality.rate` on the covariates available. **Use the settings set.seed(471) and nfolds = 10.** (name this `fit.lasso.0`).

iii) What is the `lambda.1se` value and what are the covariates in the `lambda = lambda.1se` model?

b) Lasso fit

i) Start with the `lambda = lambda.1se` model. Refit the linear model using `lm()` with the variables chosen. Perform backward elimination on this model until all features are significant at $\alpha = 0.1$ (**not 0.01**) level. Call this final model `fit.lasso`. Report the summary of `fit.lasso`.

ii) Check to see if the linear model assumptions are reasonably met for `fit.lasso`.

iii) Report the summary and explain in (non-technical) words the what do coefficients/signs of covariates in `fit.lasso` imply. In particular, add a few sentences to suggest policy makers how to lower the mortality.rate for a country?

Question 4: Prediction intervals

Lesotho is a country in Africa that is in the data set.

1) Is Lesotho's `mortality.rate` unusually high based on your final model built in `fit.lasso`? Explain why or why not. No more than three lines with quantitative support.

Part II: Logistic regression/classification, Breast Cancer Prediction

The diagnosis of breast tumors has traditionally been performed by a full biopsy, an invasive surgical procedure. Fine needle aspirations (FNAs) provide a way to examine a small amount of tissue from the tumor and the use of machine learning techniques allow classification of tumors as either benign or malignant. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Wisconsin Diagnostic Breast Cancer (WDBC) has collected data on several features of tumor cells for 569 patients.[1]

You have seen some analysis of this dataset in Quiz 3. In this exam, you will do a more refined analysis. First load the dataset using the following code.

```
# you need to put the dataset in the same folder where this .rmd file sits.
data.cancer <- read.csv("breast-cancer.csv")[,-c(1, 33)]
#names(data.cancer)
data.cancer$diagnosis <- ifelse(data.cancer$diagnosis == "M", 1, 0)
```

Question 1: Data preparation

Throughout of the remaining exam, the response will be `diagnosis`. We have coded "M" to 1 and "B" to 0.

Read the above R-Chunk carefully and answering the following questions

i) How many variables and observations does `data.cancer` have?

For our purposes, let split the data into training and test data using the following code:

```
set.seed(4712)
index.t <- sample(nrow(data.cancer), 100)
train_wdbc <- data.cancer[index.t, ]
test_wdbc  <- data.cancer[-index.t, ]
```

ii) What are the number of observations in `train_wdbc` and `test_wdbc`?

iii) What is the largest number in `index.t`?

Question 2: Linear versus Logistic Regression

In lectures you have studied linear and logistic regression but never applied them on the same dataset. Lets apply logistic and linear regression for `diagnosis` on `area_worst` using `train_wdbc`.

a) Logistic regression fit.glm

i) Report the probability equation of `P(diagnosis=1|area_worst)` using `glm()`. Call this fit `fit.glm`. (You might get a warning; ignore this) Note: no need to write a beautiful latex form. Just write down the equation in a way that we can understand it.

ii) Use .5 as the thresholing on the probability equation. Report the misclassification errors using the testing data `test_wdbc`. (Only the mis-classification error applied to the testing data is needed.)

[1] The description of the data can be found at https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names and the data is obtained from Kaggle https://www.kaggle.com/yuqing01/breast-cancer.

b) Linear regreesioin fit.lm

As another method, one could use linear regression treating `diagonosis` as a continues response variable and use the lm.fit to estimate the `P(diagnosis=1|area_worst)`.

i) Fit linear model using `lm()`, and call this fit `fit.lm`. Report the linear equation obtained.

ii) Use .5 as the thresholing on the probability this `fit.lm` equation. Report the misclassification errors using the testing data `test_wdbc`. Note: you only need to output the mis-classification error using the testing data.

c) Findings

Which method yielded a better classification rule with smaller testing misclassification error? Write a couple of sentences to comment on the fundamental differences between `fit.lm` and `fit.glm`.

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.