Midterm

STAT 471/571/701 Modern Data Mining

03/25/2019

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. 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/mortality_2012.csv`.

`/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`.

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

Variable	Description
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?

```
dim(data) # 115 observations and 30 variables
```

```
## [1] 115  30
```

ii) Are there any missing values?

```
sum(is.na(data)) # No.
```

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

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`?

```
data$Country[which.min(data$mortality.rate)]
```

```
## [1] Iceland
## 115 Levels: Algeria Argentina Armenia Australia Austria ... Vietnam
```

```
data$Country[which.max(data$mortality.rate)]
```

```
## [1] Central African Republic
## 115 Levels: Algeria Argentina Armenia Australia Austria ... Vietnam
```

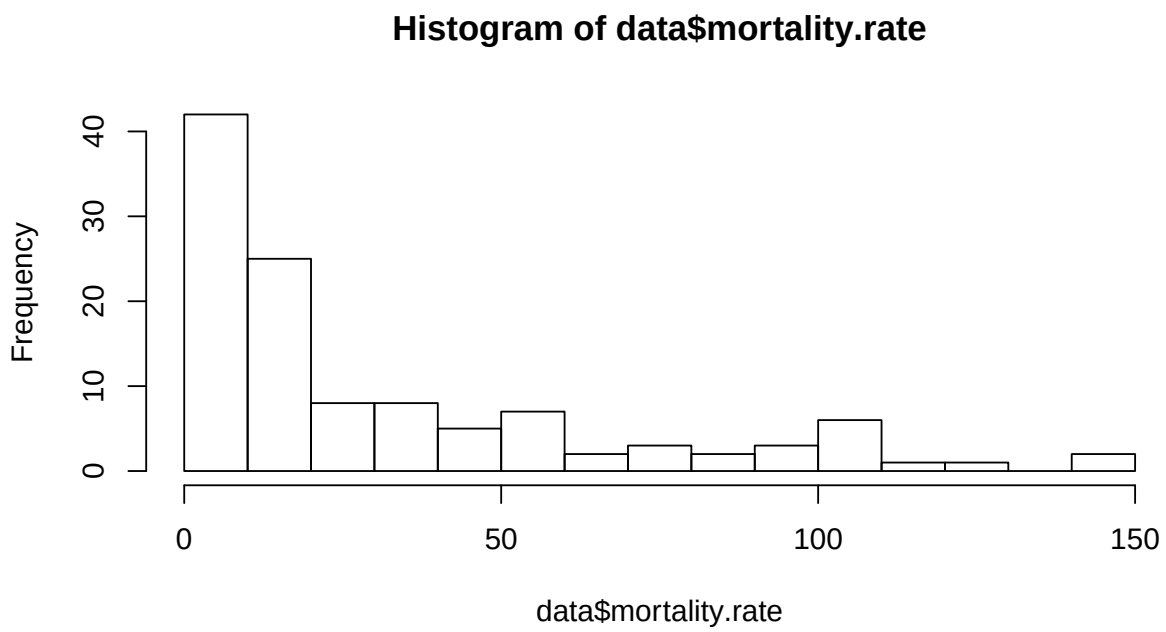
```
mean(data$mortality.rate)
```

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

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

```
## [1] 16.8
```

```
hist(data$mortality.rate, breaks=20)
```



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`?)

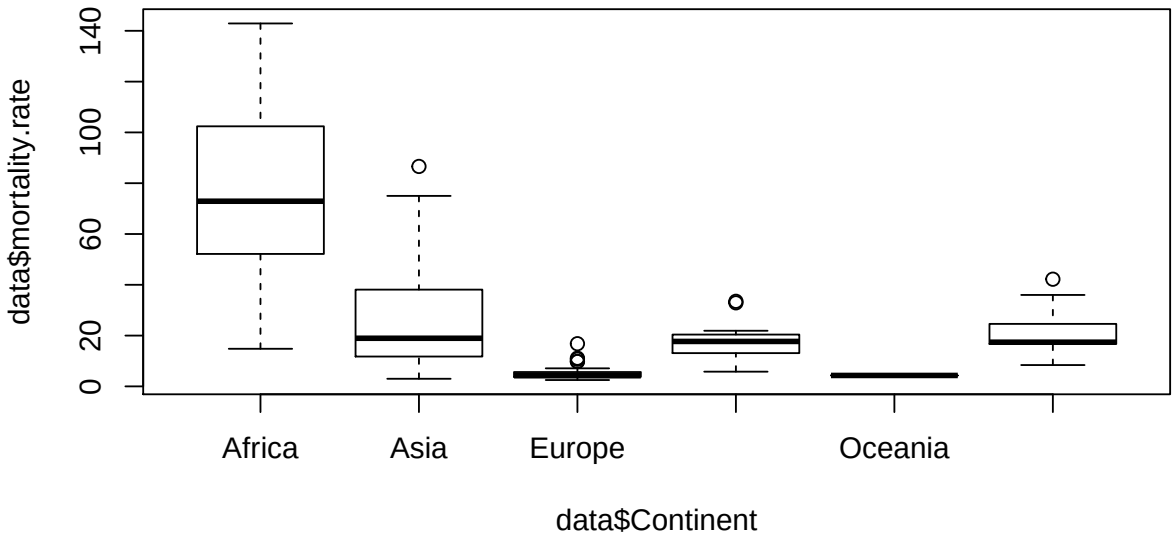
iii) Report the mean and median `mortality.rate` by `Continent`. Which `Continent` has the highest mean `mortality.rate` and what is the value?

```
data %>% group_by(Continent) %>%
  summarise(mean(mortality.rate), median(mortality.rate))
```

```
## # A tibble: 6 x 3
##   Continent     `mean(mortality.rate)` `median(mortality.rate)`
##   <fct>                          <dbl>                    <dbl>
## 1 Africa                         75.4                      72.9
## 2 Asia                           25.9                      19.0
## 3 Europe                          5.36                      4.2
## 4 North America                  18.1                      17.7
## 5 Oceania                         4.3                       4.3
## 6 South America                  21.6                      17.4
```

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

```
plot(data$mortality.rate ~ data$Continent)
```




```
data %>% group_by(Continent) %>% summarise(n())
```

```
## # A tibble: 6 x 2
##   Continent     `n()`
##   <fct>         <int>
## 1 Africa           31
## 2 Asia             28
## 3 Europe           35
## 4 North America    11
## 5 Oceania           1
## 6 South America     9
```

Question 2: Relation between mortality.rate and other variables

a) Single most usueful factor

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?

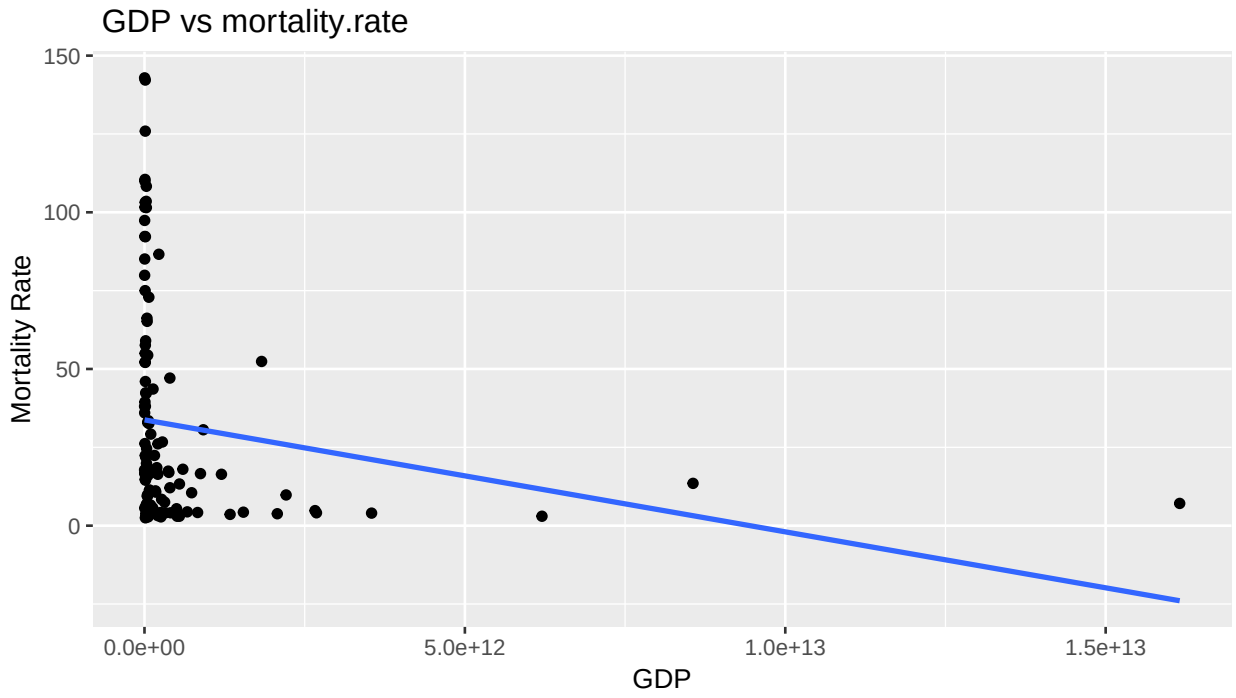
iii) Which is a better model choice? And why? No more than three sentences.

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`.

```
# i)
summary(lm(mortality.rate~GDP, data))
```

```
##
## Call:
## lm(formula = mortality.rate ~ GDP, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.21 -25.26 -15.03  13.53 109.15
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.376e+01  3.365e+00   10.03   <2e-16 ***
## GDP         -3.573e-12  1.751e-12   -2.04   0.0436 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.47 on 113 degrees of freedom
## Multiple R-squared:  0.03553,    Adjusted R-squared:  0.027
## F-statistic: 4.163 on 1 and 113 DF,  p-value: 0.04365
```

```
ggplot(data,aes(x=GDP, y=mortality.rate)) + geom_point() +
  geom_smooth(method = 'lm', se=F) +
  labs( title =" GDP vs mortality.rate", x = "GDP", y = "Mortality Rate")
```

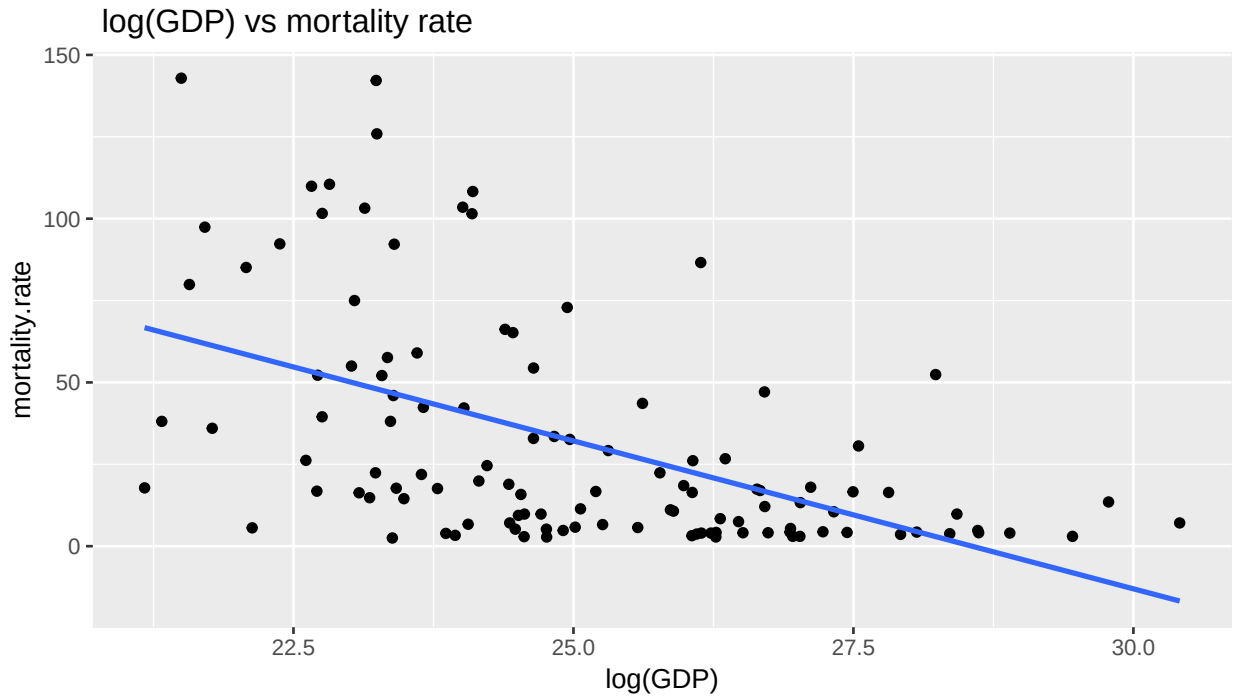



```
# ii)
summary(lm(mortality.rate~log(GDP), data))
```

```
##
```

```
## Call:
## lm(formula = mortality.rate ~ log(GDP), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.449 -20.913  -4.605  10.785  94.151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  257.941     34.940   7.382 2.84e-11 ***
## log(GDP)      -9.032      1.391  -6.495 2.31e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.95 on 113 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2654
## F-statistic: 42.19 on 1 and 113 DF,  p-value: 2.308e-09
```

```
ggplot(data,aes(x=log(GDP), y=mortality.rate)) + geom_point() +
  geom_smooth(method = 'lm', se=F)  +
  labs( title =" log(GDP) vs mortality rate", x = "log(GDP)", y = "mortality.rate")
```




```
# iii)
# The mortality.rate vs. log(GDP) is better.
# The scatter plot of mortality.rate vs. GDP shows the relationship between the two is clearly nonlinea

# iv)
# log(GDP) is significant at the .01 level. For a single unit change in log(GDP), on average, mortality
```

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.

```
# log GDP
data$log.gdp <- log(data$GDP)
# take away the original GDP
fit2 <- lm(mortality.rate~log.gdp+Sanitation.faci+Continent, data)
summary(fit2)
```

```
##
## Call:
## lm(formula = mortality.rate ~ log.gdp + Sanitation.faci + Continent,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.705  -7.486  -0.411   5.359  51.632
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            139.44817   19.30426   7.224 7.79e-11 ***
## log.gdp                 -1.46666    0.83345  -1.760 0.081309 .
## Sanitation.faci         -0.77443    0.08033  -9.640 3.33e-16 ***
## ContinentAsia          -15.38648    5.10206  -3.016 0.003202 **
## ContinentEurope        -22.32766    5.85001  -3.817 0.000227 ***
## ContinentNorth America -20.44936    6.39035  -3.200 0.001809 **
## ContinentOceania       -16.54344   16.38432  -1.010 0.314911
## ContinentSouth America -16.86861    6.73657  -2.504 0.013788 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.24 on 107 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.8097
## F-statistic: 70.31 on 7 and 107 DF,  p-value: < 2.2e-16
```

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

```
# No. Its p-value is 0.081309 larger than 0.01.
```

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

```
# This means test all Continent coefficients are zero. Remember the base is Africa so testing coefficie
Anova(fit2)
```

```
## Anova Table (Type II tests)
##
## Response: mortality.rate
##                   Sum Sq  Df F value    Pr(>F)
## log.gdp            719.4   1  3.0967   0.08131 .
## Sanitation.faci  21590.5   1 92.9325 3.333e-16 ***
## Continent         3681.4   5  3.1692   0.01044 *
## Residuals        24858.7 107
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# the p-value is 0.01044 less then 0.05 and hence we reject the null hypothesis that all coefficients a
# iii) No. From the Anova table, we have strong evidence to reject the null hypotheses of all the conti
```

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.)

```
# iv) Africa which is the base level of Continent.
```

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.)

```
data1 <- data %>% dplyr::select(-GDP, -Country)
#str(data1)
```

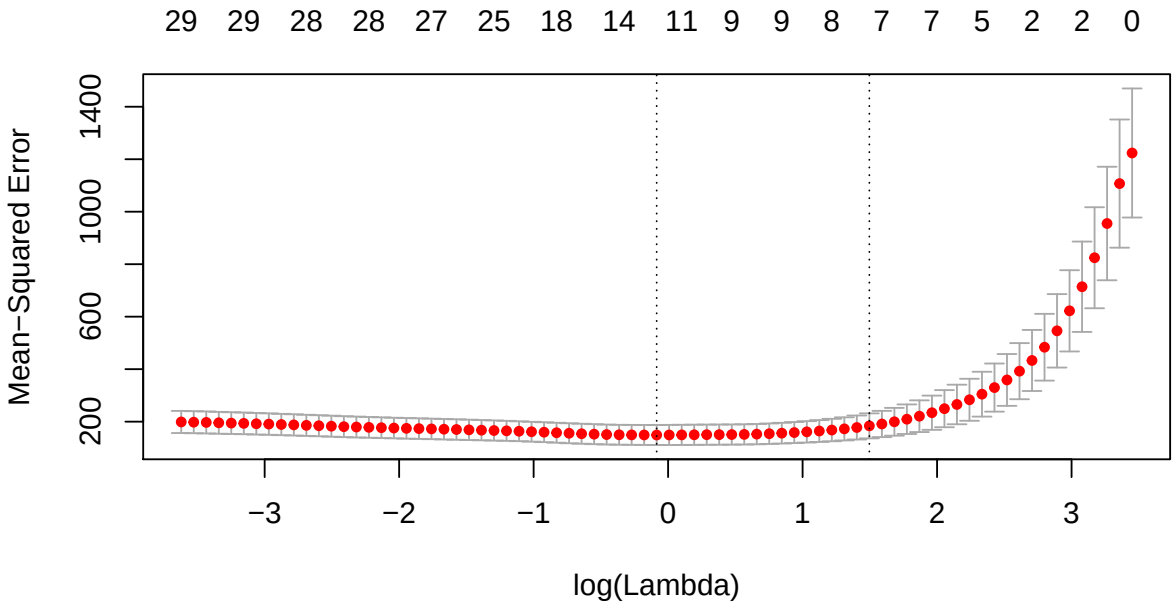
a) LASSO Regression: fit.lasso.0

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

ii) LASSO Regression: fit.lasso.0

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).

```
set.seed(471)
X <- model.matrix(mortality.rate~.,data=data1)[,-1]
Y <- data1$mortality.rate
fit.lasso.0 <- cv.glmnet(X,Y, nfolds = 10)
plot(fit.lasso.0)
```




```
# i)
# We cannot include Country because the data does not vary at the country level and as a result, a mode
```

iii) What are the `lambda.min` and `lambda.1se` values? What are the covariates in `lambda = lambda.min` and `lambda = lambda.1se` models?

```
# ii)
fit.lasso.0$lambda.min
```

```
## [1] 0.9184784
```

```
# It represents value of lambda that gives minimum cvm. In this process, it thus gives us the lambda th
# ii)
coef.min <- coef(fit.lasso.0, s="lambda.min")
coef.min <- coef.min[which(coef.min != 0),]
coef.min
```

```
##                (Intercept)  adolescent.fertility.rate
##                96.47471993                 0.03244613
## agri.forestry.fish.gdp.pct            fertility.rate
##                 0.20832432                 9.98742411
##               unemployment    ContinentNorth America
##                 0.07121506                -1.89683430
##                  Urban.pop              Immunization
##                -0.09686141                -0.28225158
##           Sanitation.faci      Immunization.measles
##                -0.23284547                -0.47228594
##                Mobile.cel            Internet.users
##                -0.01970059                -0.07917698
```

```
coef.1se <- coef(fit.lasso.0, s="lambda.1se")
coef.1se <- coef.1se[which(coef.1se != 0),]
coef.1se
```

```
##                (Intercept)  adolescent.fertility.rate
##                72.24978747                 0.01281052
## agri.forestry.fish.gdp.pct            fertility.rate
##                 0.24144175                 9.52297090
##               Immunization           Sanitation.faci
##                -0.18440515                -0.27425851
##       Immunization.measles            Internet.users
##                -0.34044539                -0.05372153
```

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`.

```
summary(lm(mortality.rate ~ adolescent.fertility.rate+agri.forestry.fish.gdp.pct+fertility.rate+Immuniza
```

```
##
## Call:
## lm(formula = mortality.rate ~ adolescent.fertility.rate + agri.forestry.fish.gdp.pct +
##     fertility.rate + Immunization + Sanitation.faci + Immunization.measles +
##     Internet.users, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.852  -5.164   0.009   4.904  48.309
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 94.35444   16.43733   5.740 8.92e-08 ***
## adolescent.fertility.rate    0.02834    0.05021   0.564  0.57367
## agri.forestry.fish.gdp.pct   0.34887    0.15324   2.277  0.02480 *
## fertility.rate               9.89152    1.50365   6.578 1.80e-09 ***
## Immunization                -0.26156    0.25420  -1.029  0.30582
## Sanitation.faci             -0.24752    0.07880  -3.141  0.00218 **
## Immunization.measles        -0.53056    0.23996  -2.211  0.02916 *
## Internet.users              -0.12529    0.06194  -2.023  0.04560 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.48 on 107 degrees of freedom
## Multiple R-squared:  0.9156, Adjusted R-squared:   0.91
## F-statistic: 165.7 on 7 and 107 DF,  p-value: < 2.2e-16
```

```
summary(lm(mortality.rate ~ agri.forestry.fish.gdp.pct+fertility.rate+Immunization+Sanitation.faci+Immu
```

```
##
## Call:
## lm(formula = mortality.rate ~ agri.forestry.fish.gdp.pct + fertility.rate +
##     Immunization + Sanitation.faci + Immunization.measles + Internet.users,
##     data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.484  -5.297   0.106   4.923  48.580
```

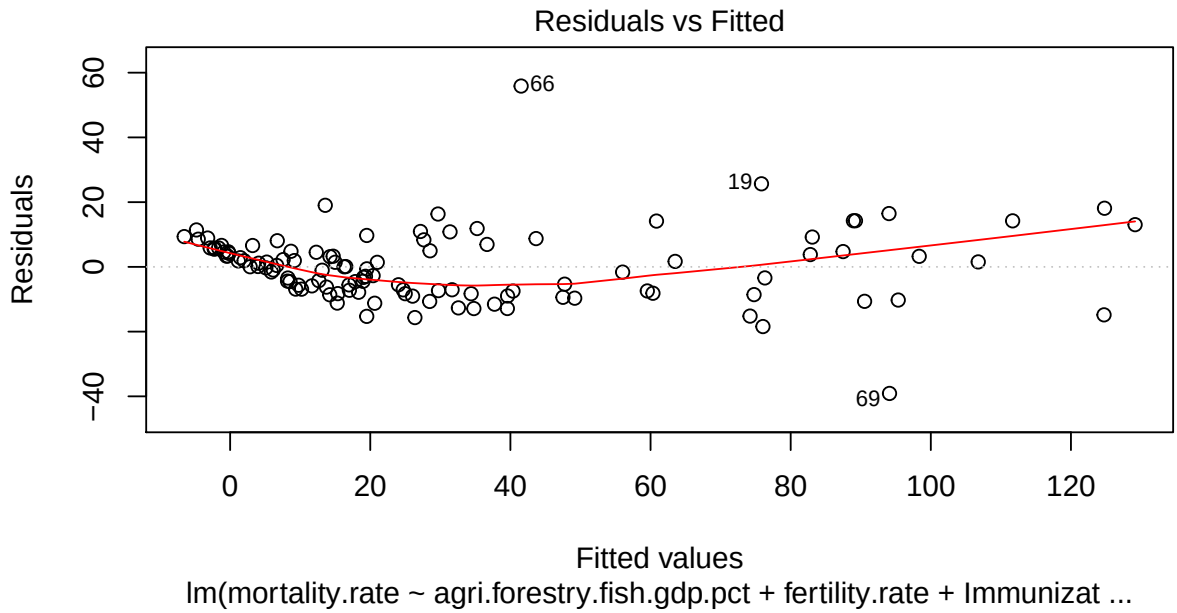
```
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 98.60519   14.56410   6.770 6.94e-10 ***
## agri.forestry.fish.gdp.pct  0.33296    0.15015   2.218 0.028683 *
## fertility.rate              10.23126    1.37354   7.449 2.45e-11 ***
## Immunization                -0.33067    0.22206  -1.489 0.139383
## Sanitation.faci             -0.25930    0.07575  -3.423 0.000875 ***
## Immunization.measles        -0.48661    0.22626  -2.151 0.033730 *
## Internet.users              -0.13548    0.05906  -2.294 0.023729 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.45 on 108 degrees of freedom
## Multiple R-squared:  0.9153, Adjusted R-squared:  0.9106
## F-statistic: 194.5 on 6 and 108 DF,  p-value: < 2.2e-16
```

```
fit.lasso <- lm(mortality.rate ~ agri.forestry.fish.gdp.pct+fertility.rate+Immunization.measles+Interne
summary(fit.lasso)
```

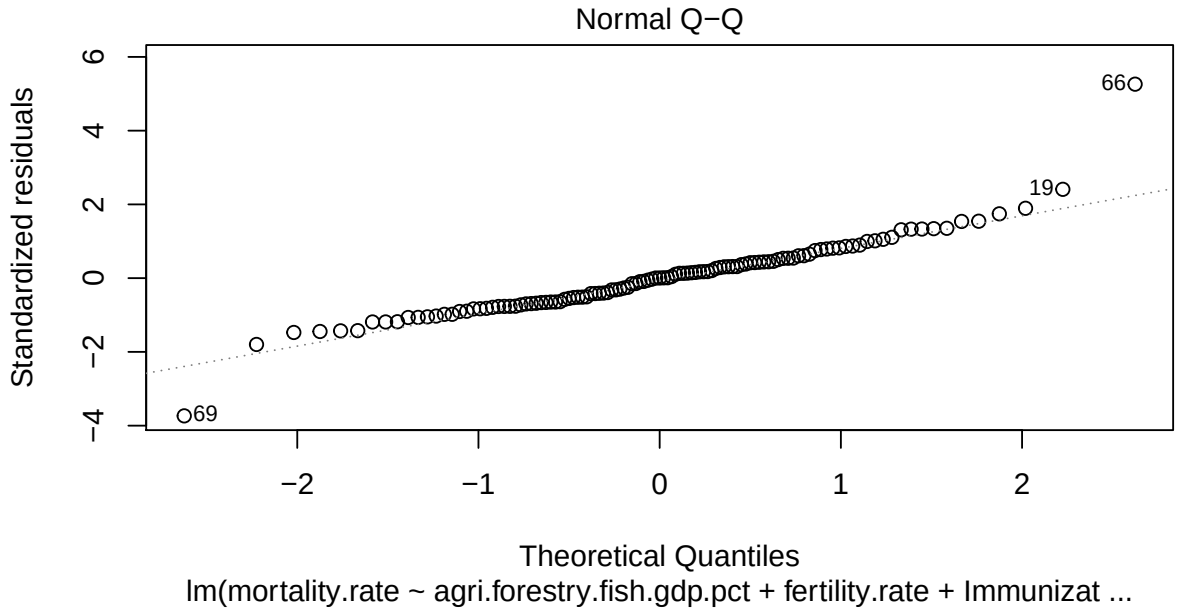
```
##
## Call:
## lm(formula = mortality.rate ~ agri.forestry.fish.gdp.pct + fertility.rate +
##     Immunization.measles + Internet.users, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.105  -7.278   0.046   5.648  55.870
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 81.40380   14.12379   5.764 7.62e-08 ***
## agri.forestry.fish.gdp.pct  0.46768    0.15270   3.063 0.002757 **
## fertility.rate              12.46233    1.28896   9.669 2.25e-16 ***
## Immunization.measles        -0.88879    0.13224  -6.721 8.33e-10 ***
## Internet.users              -0.21424    0.05651  -3.792 0.000245 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.97 on 110 degrees of freedom
## Multiple R-squared:  0.9048, Adjusted R-squared:  0.9014
## F-statistic: 261.5 on 4 and 110 DF,  p-value: < 2.2e-16
```

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

```
plot(fit.lasso, 1)
```




```
plot(fit.lasso, 2)
```




```
data1[66, ]
```

```
##    mortality.rate adolescent.fertility.rate agri.forestry.fish.gdp.pct
## 66           97.4                    90.491                   4.999449
##    industry.gdp.pct      CO2 fertility.rate GDP.per.capita GDP.grwoth.rate
## 66         29.16189 1.151021          3.253       1281.612        5.998072
```

```
##           GNI inflation population.growth population unemployment
## 66 6433653791  2.481441          1.240334    2089928       23.157
##    Continent Urban.pop Household.consump Forest.area Arable.land Water
## 66    Africa     25.76          7.719094        1.52        0.14  81.2
##    Food.prod.index Health.expend Immunization Sanitation.faci
## 66           98.65          9.55           96            29.3
##    Immunization.measles Health.exp.pocket Fixed.tel Mobile.cel
## 66                   92             14.77      2.47       75.3
##    Internet.users  log.gdp
## 66           4.59 21.70851
```

```
data[66,]
```

```
##    Country mortality.rate adolescent.fertility.rate
## 66 Lesotho           97.4                    90.491
##    agri.forestry.fish.gdp.pct industry.gdp.pct      CO2 fertility.rate
## 66                   4.999449         29.16189 1.151021          3.253
##           GDP GDP.per.capita GDP.grwoth.rate        GNI inflation
## 66 2678475775       1281.612        5.998072 6433653791  2.481441
##    population.growth population unemployment Continent Urban.pop
## 66          1.240334    2089928       23.157    Africa     25.76
##    Household.consump Forest.area Arable.land Water Food.prod.index
## 66          7.719094        1.52        0.14  81.2           98.65
##    Health.expend Immunization Sanitation.faci Immunization.measles
## 66          9.55           96            29.3                   92
##    Health.exp.pocket Fixed.tel Mobile.cel Internet.users  log.gdp
## 66             14.77      2.47       75.3           4.59 21.70851
```

iii) Report the summary and explain in (non-technical) words the what do coefficients/signs of covariates in `fit.lasso` imply. Inparticular, 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) Provide a 95% prediction interval of the `mortality.rate` for Lesotho.

2) Is Lesotho's mortality.rate unusually high? Explain why or why not.

```
predict(fit.lasso, data[66,],interval ="prediction", se.fit = TRUE)
```

```
## $fit
##         fit      lwr      upr
## 66 41.52962 19.10539 63.95385
##
## $se.fit
## [1] 2.760257
##
## $df
## [1] 110
##
## $residual.scale
## [1] 10.97345
```

```
data[66, ]
```

```
##    Country mortality.rate adolescent.fertility.rate
## 66 Lesotho           97.4                    90.491
##    agri.forestry.fish.gdp.pct industry.gdp.pct      CO2 fertility.rate
## 66                   4.999449         29.16189 1.151021          3.253
##           GDP GDP.per.capita GDP.grwoth.rate        GNI inflation
## 66 2678475775       1281.612        5.998072 6433653791  2.481441
##    population.growth population unemployment Continent Urban.pop
## 66          1.240334    2089928       23.157    Africa     25.76
##    Household.consump Forest.area Arable.land Water Food.prod.index
## 66          7.719094        1.52        0.14  81.2           98.65
##    Health.expend Immunization Sanitation.faci Immunization.measles
## 66          9.55           96            29.3                  92
##    Health.exp.pocket Fixed.tel Mobile.cel Internet.users  log.gdp
## 66             14.77      2.47       75.3           4.59 21.70851
```

Part II: 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)
```

```
##  [1] "diagnosis"              "radius_mean"
##  [3] "texture_mean"           "perimeter_mean"
##  [5] "area_mean"              "smoothness_mean"
##  [7] "compactness_mean"       "concavity_mean"
##  [9] "concave.points_mean"    "symmetry_mean"
## [11] "fractal_dimension_mean" "radius_se"
## [13] "texture_se"             "perimeter_se"
## [15] "area_se"                "smoothness_se"
## [17] "compactness_se"         "concavity_se"
## [19] "concave.points_se"      "symmetry_se"
## [21] "fractal_dimension_se"   "radius_worst"
## [23] "texture_worst"          "perimeter_worst"
## [25] "area_worst"             "smoothness_worst"
## [27] "compactness_worst"      "concavity_worst"
## [29] "concave.points_worst"   "symmetry_worst"
## [31] "fractal_dimension_worst"
```

```
data.cancer$diagnosis <- ifelse(data.cancer$diagnosis == "M", 1, 0)
#data.cancer$diagnosis <- as.factor(data.cancer$diagnosis)
# str(data.cancer)
# split data into training and testing sets
set.seed(4712)
index.t <- sample(nrow(data.cancer), 100)
train_wdbc <- data.cancer[index.t, ]
test_wdbc  <- data.cancer[-index.t, ]
```

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 follwing questions

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

[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.

```r
dim(data.cancer)
```

```
## [1] 569  31
```

```r
# 569 observations and 31 variables.
```

ii) How many of the observations are there for `train_wdbc` and `test_wdbc`?

```r
dim(train_wdbc) # 100 observations and 31 variables
```

```
## [1] 100  31
```

```r
dim(test_wdbc) # 469 observations and 31 variables
```

```
## [1] 469  31
```

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

```r
max(index.t) # 556.
```

```
## [1] 556
```

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 probabilty equation of `P(diagnosis=1|area_worst)` using `glm()`. Call this fit `fit.glm`.

```r
# logistic regression of diagnosis on area_worst.
fit.glm <- glm(diagnosis ~ area_worst, train_wdbc, family = binomial(logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
fit.glm$coefficients
```

```
##  (Intercept)   area_worst
## -8.110600271  0.009214652
```

```r
# Equation: P(diag = 1|area_worst) = exp(-8.110600271 + 0.009214652*area_worst)/(1 + exp(...))
```

ii) Use .5 as the thresholing on the probability equation. Report the misclassification errors using the testing data `test_wdbc`.

```r
fit.glm.pred <- predict(fit.glm, test_wdbc, type = "response") >= 0.5
mean(fit.glm.pred != test_wdbc$diagnosis)
```

```
## [1] 0.07462687
```

```r
# misclassification error is 0.07462687.
```

b) Linear regreesioin fit.lm

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

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

```r
# linear regression of diagnosis on area_worst.
fit.lm <- lm(diagnosis ~  area_worst, train_wdbc)
fit.lm$coefficients
```

```
##   (Intercept)    area_worst
## -0.0628137600  0.0004569056
```

```r
# Equation: diagnosis = -0.0628137600 + 0.0004569056
```

ii) Use .5 as the thresholing on the probability this fit.lm equation. Report the misclassification errors using the testing data `test_wdbc`.

```r
fit.lm.pred <- predict(fit.lm, test_wdbc) >= 0.5
mean(fit.lm.pred != test_wdbc$diagnosis)
```

```
## [1] 0.1599147
```

```r
# misclassification error is 0.1599147.
```

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.

```r
# fit.glm yielded a better result than fit.lm which is probably expected in that fit.glm respects the f
```

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.