Midterm: COVID-19 Case Study, solution

Linda Zhao

March 27, 2023

Contents

Instruction

All the teaching team members will be available from 7:00 - 9:10 PM. The submission will be closed sharp at 9:10PM.

Instruction: This midterm requires you to use R. It is completely open book/notes/internet. However any versions of ChatGPT or alike are strictly prohibitive. 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 which is the default setup for this file. 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 the data using R. Make sure the compiled html (and/or pdf) file 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. **Whenever we ask for test at some significant level, assume all the model assumptions are satisfied.**

All the answers should be clearly supported by relevant R code or based on the R output.

There are 4 questions with various parts:

- **Question 1: 3 parts**
- **Question 2: 3 parts**
- **Question 3: 2 parts**
- **Question 4: 4 parts**

DO NOT spend too much time on a single question. Come back to where you stuck after you have tried all the questions.

Data needed for the Midterm: - Canvas → Files → Exams → Midterm → Midterm Spring 2023 → midterm.csv. - Place the dataset `midterm.csv` into the same directory of the this Rmarkdown file.

Electronic Submission: Two files needed: your `.rmd` file and a compiled html file.

Label them with your full name. In the `Assignments` section, go to the `Midterm` assignment and upload your completed files. If you have trouble submitting the files to Canvas, email them to lzhao@wharton.upenn.edu and dongwooo@wharton.upenn.edu

The submission folder will be closed sharp at 9:10PM.

On-site Help: We will answer any clarification questions. We may also help out with some minor code issues. We will not, however, provide any answers as to what functions to use for example.

Raise your hand if you want to talk to one of us.

In case of emergency, here is Linda's cell: 6106590187 (text or call her)

Background

The outbreak of the novel Corona virus disease 2019 (COVID-19) was declared a public health emergency of international concern by the World Health Organization (WHO) on January 30, 2020. Upwards of 112 million cases have been confirmed worldwide, with nearly 2.5 million associated deaths. Within the US alone, there have been over 500,000 deaths and upwards of 28 million cases reported. Governments around the world have implemented and suggested a number of policies to lessen the spread of the pandemic, including mask-wearing requirements, travel restrictions, business and school closures, and even stay-at-home orders. The global pandemic has impacted the lives of individuals in countless ways, and though many countries have begun vaccinating individuals, the long-term impact of the virus remains unclear.

The impact of COVID-19 on a given segment of the population appears to vary drastically based on the socioeconomic characteristics of the segment. In particular, differing rates of infection and fatalities have been reported among different racial groups, age groups, and socioeconomic groups. One of the most important metrics for determining the impact of the pandemic is the death rate, which is the proportion of people within the total population that die due to the the disease.

There are two main goals for this case study.

1. Number of deaths vary drastically across US regions. US regions can help to describe a larger area and also helps to group together states that are similar in features such as geography, culture, history, and climate. We want to find out how regions relate to the fatality rate.
2. There have been studies on COVID income disparities. Is there evidence in our data to show that the level of household income relates to the death at county level?

To make our case study here simple and manageable in a timely fashion, we have assembled a subset of data called: `midterm.csv`. It includes county level total number of deaths by a chosen date, together with

selected demographic information. `State` names and `Region` are also included in the data. The name of each variable should be self-explanatory.

Data preparation

In this case study, we have created `midterm.csv` based on the following **THREE** cleaned datasets:

- **covid_county.csv**: County-level socioeconomic information that combines the above-mentioned 4 datasets: Income (Poverty level and household income), Jobs (Employment type, rate, and change), People (Population size, density, education level, race, age, household size, and migration rates), County Classifications
- **covid_rates.csv**: Daily cumulative numbers on infection and fatality for each county
- **state_region.csv**: Grouping states into distinct regions in the United States

Among all data, the unique identifier of county is `FIPS`.

What is a good way to measure COVID death rate? There are quite a number of counties with a very low or zero number of deaths. We have proposed an effective way of handling such an imbalanced situation. In the following chunk, we created a new measurement of fatality rate

$$\text{fatality\_rate} = \frac{\text{cum\_deaths} + 1}{\text{cum\_cases} + 2}$$

and then applied the log function to get the variable labeled as `log_fatality_rate`.

Not only that, we will focus on the median household income `MedHHInc_10k` whose unit of measurement is \$10,000. This is exactly equal to `MedHHInc` divided by 10,000.

We then created the data `midterm.csv` by combining the `log_fatality_rate` with a subset of county level demographic information. The process of creating this data is shown in the following R-chunk, labeled as `data prep`. Read through the `data prep` chunk carefully for future analysis.

```
# This is how we created the midterm.csv
# DONOT RUN this chunk

# county-level socialeconomic information
covid_county <- read_csv("covid_county.csv")
# county-level COVID case and death
covid_rates <- read_csv("covid_rates.csv")
# region dataset
state_region <- read_csv("state_region.csv")

dat <- covid_rates %>%
  #filter(date == "2020-09-01") %>%
  filter(date == "2020-12-31") %>%
  mutate(log_fatality_rate = log((cum_deaths + 1) / (cum_cases + 2))) %>%
  select(FIPS, log_fatality_rate)

# join with county-level demographic data
dat <- left_join(dat, covid_county, by = "FIPS") %>%
    left_join(state_region, by = "State") %>%
    mutate(MedHHInc_10k = MedHHInc / 10000) %>%
    drop_na()

# take a subset of the demographic info
dat <- dat %>%
    select(log_fatality_rate, MedHHInc_10k, Region, State, County, UnempRate2019, PctEmpFIRE, PctEmpCons
```

```
# output
write_csv(dat, "midterm.csv")
```

We next read the pre-processed data `midterm.csv` into R and label it as `dat`.

`midterm.csv` must be stored in the same directory of this Rmd file.

`dat` will be used throughout the midterm.

```
dat <- read.csv("midterm.csv")
#summary(dat)
```

Question 1. EDA

We first study the log fatality rates at region and household-income level via the following EDA.

Question 1-1.

Report the following summary statistics:

a. How many Regions are there?

b. How many Counties are there in each Region?

Answer:

```
dat %>%
  group_by(Region) %>%
  summarise(num_County = n())
```

```
## # A tibble: 5 x 2
##   Region    num_County
##   <chr>          <int>
## 1 Midwest         1006
## 2 Northeast        244
## 3 Southeast       1057
## 4 Southwest        360
## 5 West             344
```

This is a reference. Didn't do it.

```
dat %>%    # this is only for me to check something
  group_by(Region, State) %>%
  summarise(num_State = n_distinct(County))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 48 x 3
## # Groups:   Region [5]
##    Region  State num_State
##    <chr>   <chr>     <int>
##  1 Midwest IA           99
##  2 Midwest IL          100
##  3 Midwest IN           92
##  4 Midwest KS          102
##  5 Midwest MI           83
##  6 Midwest MN           87
```

```
##  7 Midwest MO          113
##  8 Midwest ND           43
##  9 Midwest NE           73
## 10 Midwest OH           88
## # ... with 38 more rows
```

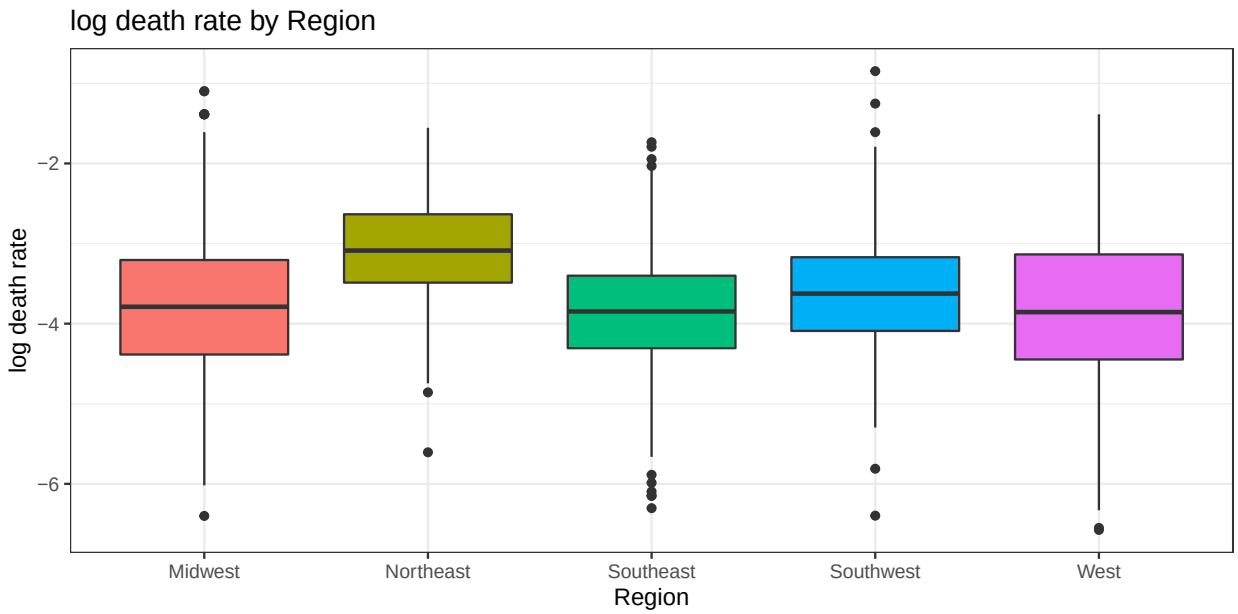
Question 1-2

a. Display back to back boxplots of `log_fatality_rate` with respect to Regions.

b. Using no more than two sentences, comment on the variations of `log_fatality_rate` among different Regions.

Answer:

```
dat %>%
  ggplot(aes(x = Region, y = log_fatality_rate, fill = Region)) +
  geom_boxplot() +
  xlab("Region") +
  ylab("log death rate") +
  ggtitle("log death rate by Region") +
  theme_bw() +
  theme(legend.position = "none")
```



The death rates are higher in NE while similar in other regions.

Question 1-3.

We would like to explore how `MedHHInc_10k` and `Region` relate to `log_fatality_rate`.

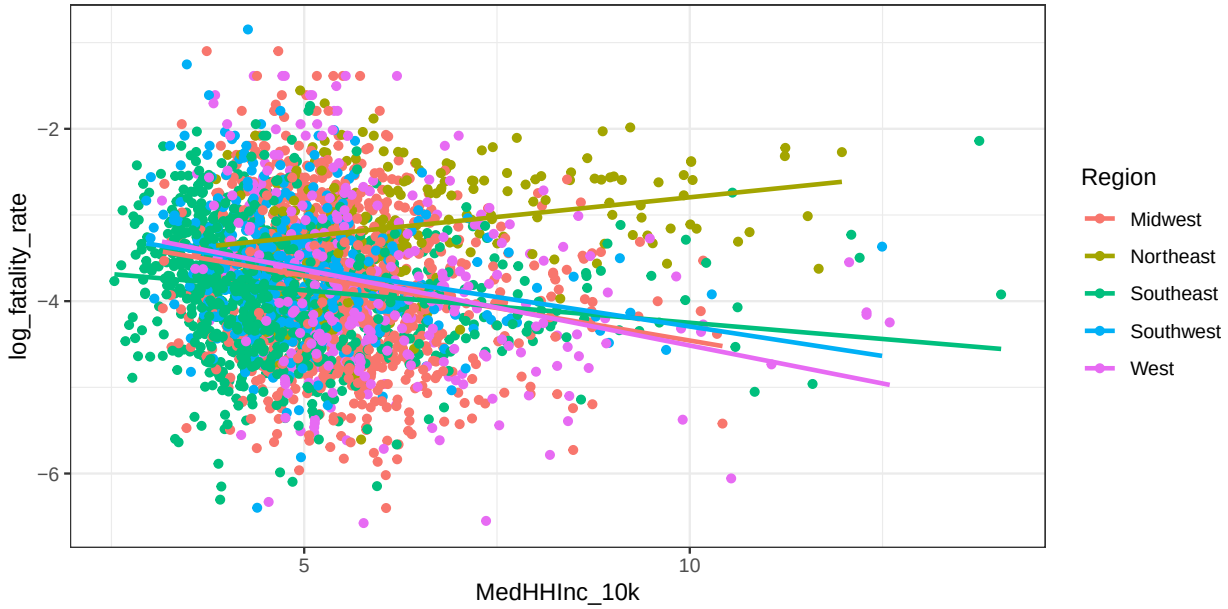
a. Draw scatter plots of `log_fatality_rate` versus `MedHHInc_10k` with a line of simple linear regression fit with respect to `Region`.
   - Option 1: You can make a single plot with five lines.
   - Option 2: Produce five plots and place them side-by-side. (Hint: `facet_wrap` function would be helpful to automatically draw many plots side-by-side. Consult with our first Module.)

- Option 3: If you have trouble to produce either one of the above just provide plots one by one or only provide plots for `Northeast` and `Southeast` region.

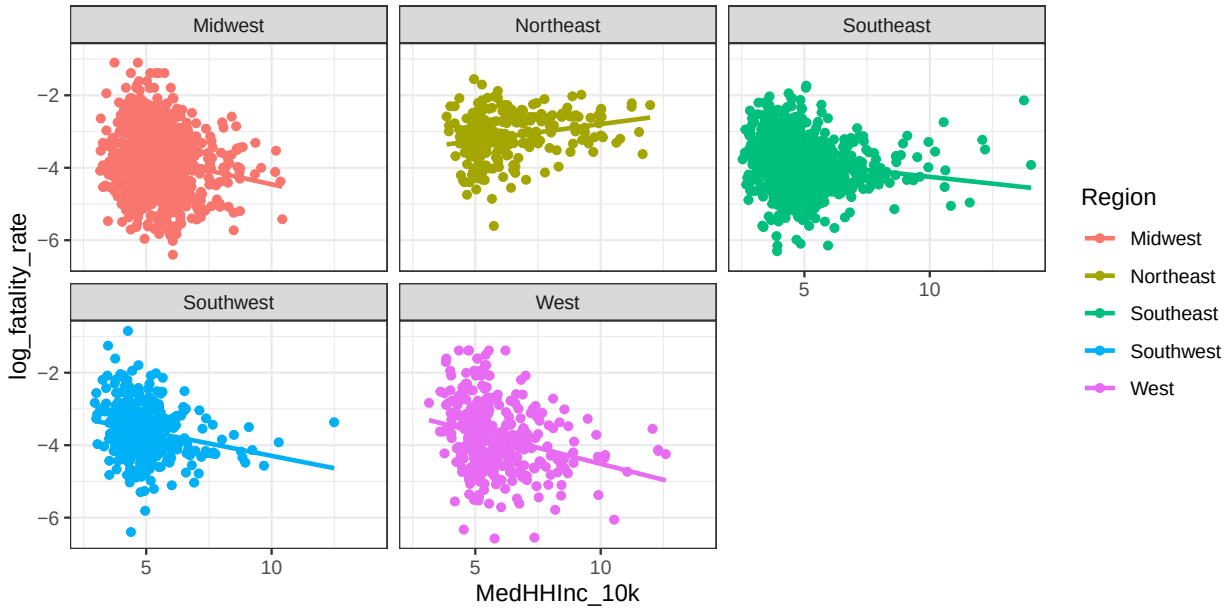
b. Do you see possible interaction effects of `MedHHInc_10k` and `Region` over `log_fatality_rate`? How so? Use no more than two sentences to explain.

Answer: Yes, since the slope for `Northeast` is positive while all others are negative.

```
dat %>%
  ggplot(aes(x=MedHHInc_10k, y=log_fatality_rate, group = Region, color=Region)) +
  geom_point()+
  geom_smooth(method="lm", formula=y~x, se=F)+
  theme_bw()
```




```
dat %>%
  ggplot(aes(x=MedHHInc_10k, y=log_fatality_rate, group = Region, color=Region)) +
  geom_point()+
  geom_smooth(method="lm", formula=y~x, se=F)+
  facet_wrap(~Region) +
  theme_bw()
```



Answer: Since the slopes are very different among five regioins, it indicates some interaction effects.

Question 2. Simple analysis

There are a number of studies indicating that COVID is affected by income level. In the following analyses, we focus on the effect of `MedHHInc_10k` and `Region` over `log_fatality_rate`.

Question 2-1.

Run a regression of `log_fatality_rate` vs. `MedHHInc_10k` controlling `Region` (without interactions) as `fit1`. Report the summary of `fit1`. Also provide the appropriate p-values either from `fit1` or other appropriate tests to answer the following questions.

a. Can you reject the null hypothesis that `MedHHInc_10k` is not needed after controlling `Region` at $\alpha = .01$?
b. Do we have evidence to support the hypothesis that after controlling for `MedHHInc_10k`, the `log_fatality_rate` are different on average among different regions?
c. Controlling for `MedHHInc_10k`, which Region has highest `log_fatality_rate` on average?

Answer: Yes, we can reject the null according to the Z or F-statistic as follow. Controlling `MedHHInc_10k`, northeast region has the highest `log_fatality_rate` on average.

```
fit1 <- lm(log_fatality_rate ~ MedHHInc_10k + Region, data = dat)
summary(fit1)
```

```
##
## Call:
## lm(formula = log_fatality_rate ~ MedHHInc_10k + Region, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8435 -0.5042 -0.0042  0.5053  2.6952
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
```

```
## (Intercept)      -3.2795     0.0647  -50.69 < 0.0000000000000002 ***
## MedHHInc_10k     -0.0904     0.0109   -8.26  0.00000000000000022 ***
## RegionNortheast   0.7308     0.0561   13.02 < 0.0000000000000002 ***
## RegionSoutheast  -0.1468     0.0350   -4.19  0.00002876395670215 ***
## RegionSouthwest   0.1228     0.0479    2.56                 0.01 *
## RegionWest        0.0225     0.0485    0.46                 0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.774 on 3005 degrees of freedom
## Multiple R-squared:  0.0793, Adjusted R-squared:  0.0778
## F-statistic: 51.8 on 5 and 3005 DF,  p-value: <0.0000000000000002
```

```
Anova(fit1)
```

```
## Anova Table (Type II tests)
##
## Response: log_fatality_rate
##              Sum Sq   Df F value               Pr(>F)
## MedHHInc_10k     41    1    68.2  0.00000000000000022 ***
## Region          144    4    60.1 < 0.0000000000000002 ***
## Residuals      1801 3005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answers: from the above output, we know:

a. We reject the null hypothesis that `MedHHInc_10k` is not needed after controlling `Region` at $\alpha = .01$? Since the p-value = 0.00000000000000022 which is much smaller that .01

b. We have strong evidence to support the hypothesis that after controlling for `MedHHInc_10k`, the `log_fatality_rate` are different on average among different regions. Since the p-value is 0.0000000000000002

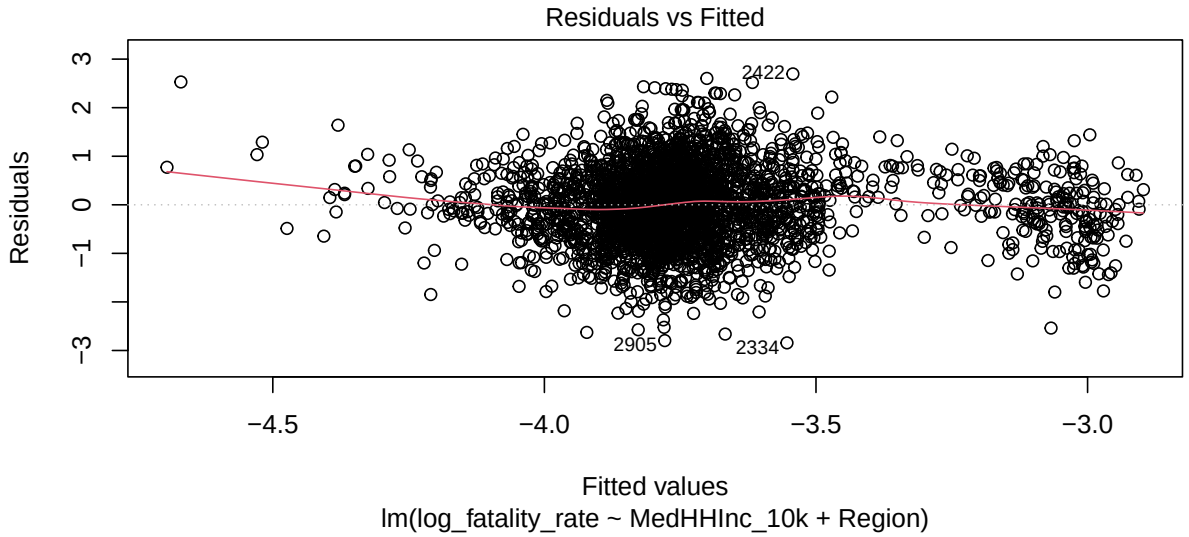
c. Controlling for `MedHHInc_10k`, RegionNortheast= 0.7308 has the highest `log_fatality_rate` on average.

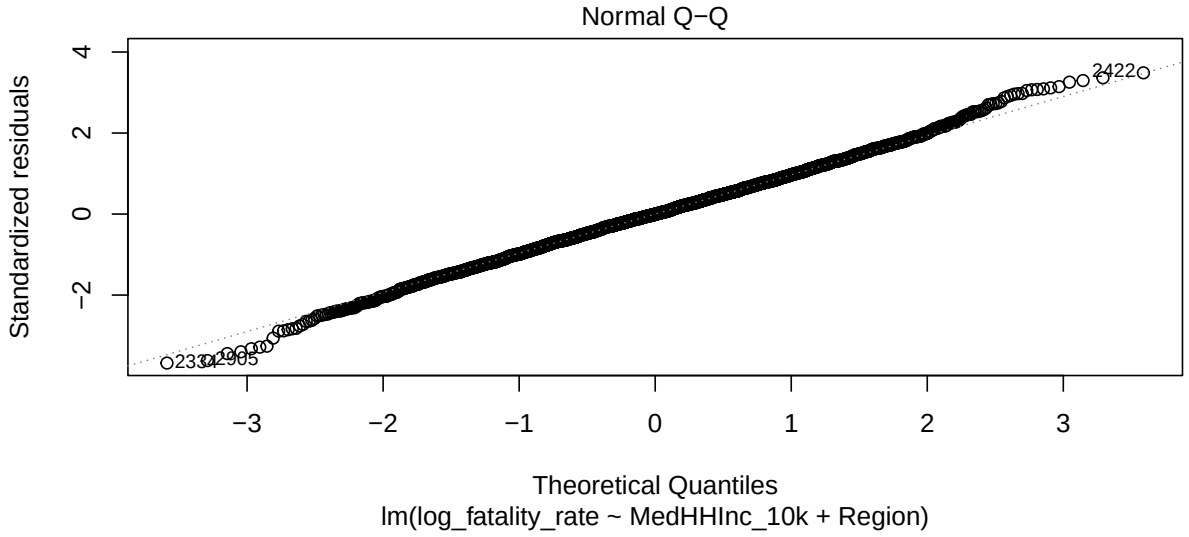
Question 2-2.

Are the linear model assumptions reasonably met in `fit1`? Provide residual and normal plots for `fit1`. Use no more than three sentences to summarize your model diagnoses.

Answer:

```
plot(fit1, 1:2)
```





Answer:

a. The linearity assumtion is fine since the residual plot show reasonable symmetry w.r.t the horizontal line at 0.

b. The qq-plot indicates the normality assumtion is fine.

Question 2-3. Prediction interval

Assume all the linear model assumptions are met in fit1. We would like to predict the fatality_rate for Philadelphia by providing a 95% prediction interval:

a. First provide a 95% prediction interval for `log_fatality_rate` here.

b. Then unlog (base e) the above 95% prediction interval to obtain 95% prediction interval of fatality rate of Philadelphia.

Hint: First extract information of Philadelphia then do prediction interval.

```
philly <- filter(dat, County == "Philadelphia")  # to extract information for philly
```

Answer:

```
philly <- filter(dat, County == "Philadelphia")  # to extract information for philly
fit <- predict(fit1, philly, interval = "predict",
        se.fit = TRUE, level = 0.95)$fit
fit
```

```
##     fit   lwr   upr
## 1 -2.97 -4.49 -1.44
```

```
fit.conf <- exp(fit)
fit.conf
```

```
##      fit    lwr   upr
## 1 0.0515 0.0113 0.236
```

Question 3. Interaction analysis

We would look into interactions between household income and region on `log_fatality_rate`.

Question 3-1.

Run a regression of `log_fatality_rate` vs. `MedHHInc_10k`, `Region` and interaction between `MedHHInc_10k` and `Region`. Save the result as `fit2`. Do we have evidence to support that there is an interaction effect of `MedHHInc_10k` and `Region` over `log_fatality_rate`? Perform a test at $\alpha = .01$.

Answer:

```
fit2 <- lm(log_fatality_rate ~ MedHHInc_10k + Region + Region * MedHHInc_10k, data = dat)
summary(fit2)
```

```
##
## Call:
## lm(formula = log_fatality_rate ~ MedHHInc_10k + Region + Region *
##     MedHHInc_10k, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8697 -0.4994  0.0009  0.4879  2.6635
##
## Coefficients:
##                               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                    -2.9482     0.1300  -22.68 < 0.0000000000000002
## MedHHInc_10k                   -0.1509     0.0233   -6.47       0.000000000116
## RegionNortheast                -0.7593     0.2313   -3.28              0.00104
## RegionSoutheast                -0.5495     0.1559   -3.52              0.00043
## RegionSouthwest                 0.0212     0.2161    0.10              0.92190
## RegionWest                      0.1956     0.2111    0.93              0.35431
## MedHHInc_10k:RegionNortheast    0.2421     0.0372    6.51       0.000000000088
## MedHHInc_10k:RegionSoutheast    0.0756     0.0292    2.59              0.00956
## MedHHInc_10k:RegionSouthwest    0.0142     0.0411    0.34              0.73045
## MedHHInc_10k:RegionWest        -0.0253     0.0360   -0.70              0.48183
##
```

```
## (Intercept)                    ***
## MedHHInc_10k                   ***
## RegionNortheast                **
## RegionSoutheast                ***
## RegionSouthwest
## RegionWest
## MedHHInc_10k:RegionNortheast ***
## MedHHInc_10k:RegionSoutheast **
## MedHHInc_10k:RegionSouthwest
## MedHHInc_10k:RegionWest
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.767 on 3001 degrees of freedom
## Multiple R-squared:  0.0969, Adjusted R-squared:  0.0942
## F-statistic: 35.8 on 9 and 3001 DF,  p-value: <0.0000000000000002
```

```
Anova(fit2)  #solution 1 or
```

```
## Anova Table (Type II tests)
##
## Response: log_fatality_rate
##                      Sum Sq   Df F value               Pr(>F)
## MedHHInc_10k             41    1    69.4 < 0.0000000000000002 ***
## Region                  144    4    61.2 < 0.0000000000000002 ***
## MedHHInc_10k:Region      34    4    14.6       0.000000000008 ***
## Residuals              1766 3001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit1, fit2) # solution 2
```

```
## Analysis of Variance Table
##
## Model 1: log_fatality_rate ~ MedHHInc_10k + Region
## Model 2: log_fatality_rate ~ MedHHInc_10k + Region + Region * MedHHInc_10k
##   Res.Df  RSS Df Sum of Sq    F         Pr(>F)
## 1   3005 1801
## 2   3001 1766  4      34.4 14.6 0.000000000008 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 3-2

Based on `fit2`, report the slopes of `MedHHInc_10k` for Region `Northeast` and `West`.

Answer:

```
fit2$coef[7] + fit2$coef[2] # Northeast
```

```
## MedHHInc_10k:RegionNortheast
##                       0.0913
```

```
fit2$coef[10] + fit2$coef[2] # West
```

```
## MedHHInc_10k:RegionWest
##                  -0.176
```

Question 4. Final analysis

Question 4-1.

Use LASSO to pick up a few variables in addition to `MedHHInc_10k` and `Region` excluding `State` and `County`.

a. `fit_lasso_cv`: Run LASSO with the following settings to get the same results.
   - Use `set.seed(12345)` to control the cross-validation
   - Use 20-fold cross validations
   - Force `MedHHInc_10k` and `Region` in all the LASSO models
   - Pick up the variables using `lambda.1se`

b. plot `fit_lasso_cv`

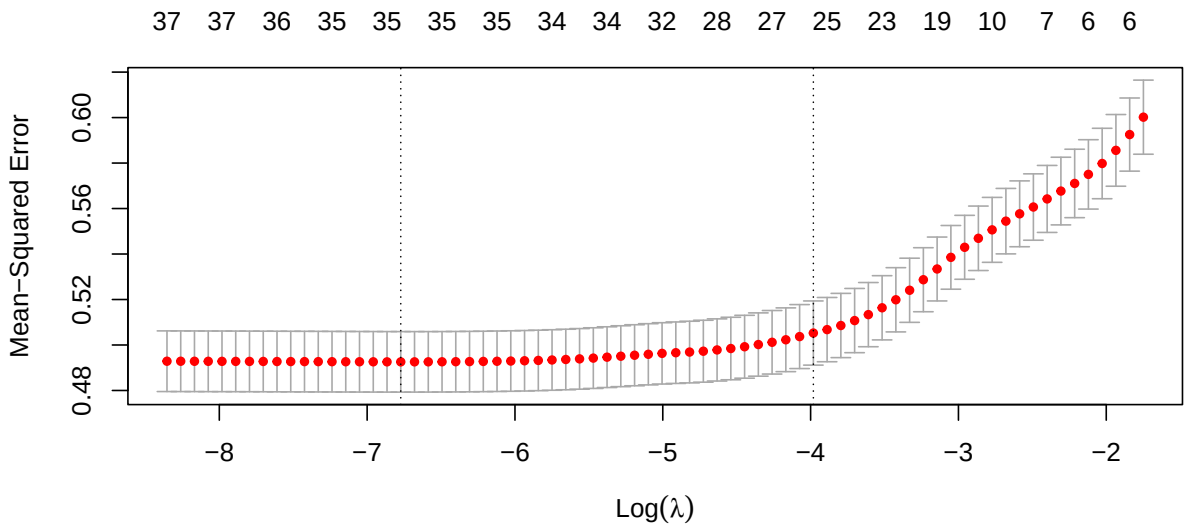
c. Save as `lasso_selected` the variable names selected by the LASSO, and print it. This vector must contain `Region` instead of each category such as `RegionWest` or `RegionSoutheast` or `RegionSouthwest`.

Hint:

```
Y <- as.matrix(dat[, 'log_fatality_rate']) # extract Y
X <- model.matrix(log_fatality_rate ~ . - State - County, data = dat)[, -1]
force_in_indicator <- c(rep(0, 5), rep(1, ncol(X)-5))
```

Answer:

```
Y <- as.matrix(dat[, 'log_fatality_rate']) # extract Y
X <- model.matrix(log_fatality_rate ~ . - State - County, data = dat)[, -1]
force_in_indicator <- c(rep(0, 5), rep(1, ncol(X)-5))
#names(dat)
set.seed(12345)
fit_lasso_cv <- cv.glmnet(X, Y, alpha = 1, nfolds = 20, intercept = TRUE,
                          penalty.factor = force_in_indicator)
plot(fit_lasso_cv)
```




```
coef_1se <- coef(fit_lasso_cv, s = "lambda.1se")
coef_1se <- coef_1se[which(coef_1se != 0), ]
lasso_selected <- c("MedHHInc_10k", "Region", names(coef_1se)[-c(1:6)])

lasso_selected
```

```
##  [1] "MedHHInc_10k"                          "Region"
##  [3] "UnempRate2019"                         "PctEmpTrans"
##  [5] "PctEmpMining"                          "PctEmpTrade"
##  [7] "PctEmpInformation"                     "PctEmpAgriculture"
##  [9] "PctEmpManufacturing"                   "PopDensity2010"
## [11] "OwnHomePct"                            "Age65AndOlderPct2010"
## [13] "TotalPop25Plus"                        "Ed4AssocDegreePct"
## [15] "ForeignBornPct"                        "NetMigrationRate1019"
## [17] "NaturalChangeRate1019"                 "WhiteNonHispanicPct2010"
## [19] "Type_2015_Update"                      "RuralUrbanContinuumCode2013"
## [21] "Perpov_1980_0711"                      "HiAmenity"
## [23] "Retirement_Destination_2015_Update"
```

Question 4-2.

Run a model `fit3` of `log_fatality_rate` vs the set of variables in `lasso_selected` given in the following code chunk. Assume this is what you have chosen using LASSO. Or those who CAN NOT run LASSO successfully can also use this set of variables to finish the `fit3`.

```
# Do not modify this code chunk
lasso_selected <- c("MedHHInc_10k", "Region", "UnempRate2019", "PctEmpTrans", "PctEmpMining", "PctEmpTra

# prepare all the variables needed in fit3
dat_fit3 <- select(dat, all_of(c("log_fatality_rate", lasso_selected)))
```

a. Report summary of `fit3`. (No need to do further model selection!)

Answering the following questions with proper p-values or perform proper tests.

a. Is `Region` significant at .01 level controlling for all other variables?

b. Is `MedHHInc_10k` significant at .01 level controlling for all other variables?

Answer: Yes.

```
#dat_fit3 <- select(dat, all_of(c("log_fatality_rate", lasso_selected)))
fit3 <- lm(log_fatality_rate ~ ., dat = dat_fit3)
# dat_fit3_1 <- dat_fit3 %>% select(-Perpov_1980_0711)
# fit3 <- lm(log_fatality_rate ~ .,  data = dat_fit3_1)
summary(fit3)
```

```
##
## Call:
## lm(formula = log_fatality_rate ~ ., data = dat_fit3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.781 -0.452  0.011  0.442  2.344
##
## Coefficients:
##                                    Estimate    Std. Error t value
## (Intercept)                     -4.8213997495  0.2273912170  -21.20
## MedHHInc_10k                     0.0615548470  0.0167630625    3.67
## RegionNortheast                  0.6849974643  0.0556313143   12.31
## RegionSoutheast                 -0.0372090373  0.0422106521   -0.88
## RegionSouthwest                  0.0367178397  0.0618861454    0.59
```

```
## RegionWest                             0.1161604829  0.0590095023    1.97
## UnempRate2019                          0.0392449340  0.0113089338    3.47
## PctEmpTrans                            0.0106806176  0.0069668936    1.53
## PctEmpMining                           0.0179557445  0.0047789027    3.76
## PctEmpTrade                           -0.0139260111  0.0055130985   -2.53
## PctEmpInformation                      0.0430821312  0.0194601500    2.21
## PctEmpAgriculture                      0.0153767394  0.0036973314    4.16
## PctEmpManufacturing                   -0.0057716910  0.0023288247   -2.48
## PopDensity2010                         0.0000271977  0.0000084033    3.24
## OwnHomePct                             0.0114663776  0.0023516245    4.88
## Age65AndOlderPct2010                   0.0508398991  0.0070199747    7.24
## TotalPop25Plus                         0.0000002352  0.0000000699    3.37
## Ed4AssocDegreePct                     -0.0233640043  0.0059925792   -3.90
## ForeignBornPct                        -0.0164379840  0.0037164835   -4.42
## NetMigrationRate1019                  -0.0062431609  0.0023877367   -2.61
## NaturalChangeRate1019                 -0.0218391096  0.0077045996   -2.83
## WhiteNonHispanicPct2010               -0.0089901026  0.0010840612   -8.29
## Type_2015_Update                      -0.0222038051  0.0075364637   -2.95
## RuralUrbanContinuumCode2013            0.0103138547  0.0073691499    1.40
## HiAmenity                             -0.1405256933  0.0397357804   -3.54
## Retirement_Destination_2015_Update -0.0840597616  0.0427255790   -1.97
##                                                    Pr(>|t|)
## (Intercept)                          < 0.0000000000000002 ***
## MedHHInc_10k                                        0.00024 ***
## RegionNortheast                      < 0.0000000000000002 ***
## RegionSoutheast                                     0.37811
## RegionSouthwest                                     0.55302
## RegionWest                                          0.04910 *
## UnempRate2019                                       0.00053 ***
## PctEmpTrans                                         0.12537
## PctEmpMining                                        0.00018 ***
## PctEmpTrade                                         0.01159 *
## PctEmpInformation                                   0.02691 *
## PctEmpAgriculture                      0.00003288314975 ***
## PctEmpManufacturing                                 0.01325 *
## PopDensity2010                                      0.00122 **
## OwnHomePct                             0.00000113957575 ***
## Age65AndOlderPct2010                   0.00000000000056 ***
## TotalPop25Plus                                      0.00077 ***
## Ed4AssocDegreePct                      0.00009878911076 ***
## ForeignBornPct                         0.00001008219191 ***
## NetMigrationRate1019                                0.00898 **
## NaturalChangeRate1019                               0.00462 **
## WhiteNonHispanicPct2010              < 0.0000000000000002 ***
## Type_2015_Update                                    0.00324 **
## RuralUrbanContinuumCode2013                         0.16174
## HiAmenity                                           0.00041 ***
## Retirement_Destination_2015_Update                  0.04923 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.699 on 2985 degrees of freedom
## Multiple R-squared:  0.254,  Adjusted R-squared:  0.247
## F-statistic: 40.6 on 25 and 2985 DF,  p-value: <0.0000000000000002
```

```
Anova(fit3)
```

```
## Anova Table (Type II tests)
##
## Response: log_fatality_rate
##                                   Sum Sq   Df F value                Pr(>F)
## MedHHInc_10k                           7    1   13.48               0.00024 ***
## Region                                87    4   44.32 < 0.0000000000000002 ***
## UnempRate2019                          6    1   12.04               0.00053 ***
## PctEmpTrans                            1    1    2.35               0.12537
## PctEmpMining                           7    1   14.12               0.00018 ***
## PctEmpTrade                            3    1    6.38               0.01159 *
## PctEmpInformation                      2    1    4.90               0.02691 *
## PctEmpAgriculture                      8    1   17.30      0.00003288314975 ***
## PctEmpManufacturing                    3    1    6.14               0.01325 *
## PopDensity2010                         5    1   10.48               0.00122 **
## OwnHomePct                            12    1   23.77      0.00000113957575 ***
## Age65AndOlderPct2010                  26    1   52.45      0.00000000000056 ***
## TotalPop25Plus                         6    1   11.33               0.00077 ***
## Ed4AssocDegreePct                      7    1   15.20      0.00009878911076 ***
## ForeignBornPct                        10    1   19.56      0.00001008219191 ***
## NetMigrationRate1019                   3    1    6.84               0.00898 **
## NaturalChangeRate1019                  4    1    8.03               0.00462 **
## WhiteNonHispanicPct2010               34    1   68.77 < 0.0000000000000002 ***
## Type_2015_Update                       4    1    8.68               0.00324 **
## RuralUrbanContinuumCode2013            1    1    1.96               0.16174
## HiAmenity                              6    1   12.51               0.00041 ***
## Retirement_Destination_2015_Update     2    1    3.87               0.04923 *
## Residuals                           1460 2985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4-3.(We took this question out of the exam.) To choose a more desirable model among `fit1` and `fit3` we recommend using C_p statistics. We can use function `AIC` to report C_p statistics as follows. `fit3` is better with a smaller C_p. Use no more than one sentence to support why C_p is a good criterion.

```
AIC(fit1)
```

```
## [1] 7011
```

```
AIC(fit3)
```

```
## [1] 6419
```

Answer: Cp estimates testing errors.

Question 4-3.

Based on `fit3`, summarize your findings focusing on the effects of `MedHHInc_10k` and `Region` over the `log_fatality_rate`. Does that make sense to you? Use no more than three sentences to describe.

Answer:

========== **End of the Midterm** ==========

Appendix

Data Summary

The data comes from several different sources:

1. County-level infection and fatality data - This dataset gives daily cumulative numbers on infection and fatality for each county.
   - NYC data
2. County-level socioeconomic data - The following are the four relevant datasets from this site.
   i. Income - Poverty level and household income.
   ii. Jobs - Employment type, rate, and change.
   iii. People - Population size, density, education level, race, age, household size, and migration rates.
   iv. County Classifications - Type of county (rural or urban on a rural-urban continuum scale).